



# **Computational Biology and High Performance Computing**

**Tutorial M4 p.m.**



**November 6, 2000  
SC'2000, Dallas, Texas**



## **Tutorial Outline**



- **8:30 a.m. - 12:00 p.m.**
  - **Introduction to Biology**
  - **Overview Computational Biology**
  - **DNA sequences**
  
- **1:30 p.m. - 5:00 p.m.**
  - **Protein Sequences**
  - **Phylogeny**
  - **Specialized Databases**

 <b>Tutorial Outline: Afternoon</b> 	
■ 1:30 p.m. - 2:00 p.m.	<b>Working with Proteins</b>
■ 2:00 p.m. - 3:00 p.m.	<b>Phylogeny</b>
■ 3:00 p.m. - 3:30 p.m.	<b>BREAK</b>
■ 3:30 p.m. - 4:30 p.m.	<b>Specialized Databases</b>
■ 4:30 p.m. - 5:00 p.m.	<b>Genetic Networks</b>
Computational Biology @ SC 2000	

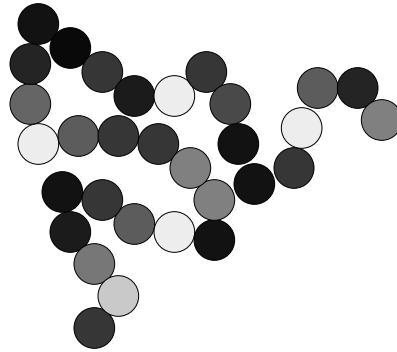


## Proteins

Manfred Zorn  
MDZorn@lbl.gov  
NERSC

# What is a protein?

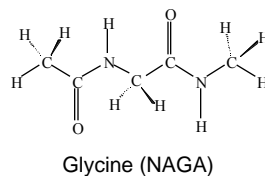
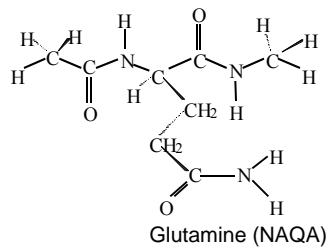
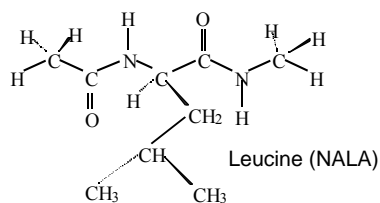
A biopolymer which is distinct from a heteropolymer in one very important way  
It's 3-D structure is uniquely tailored to perform a specific function

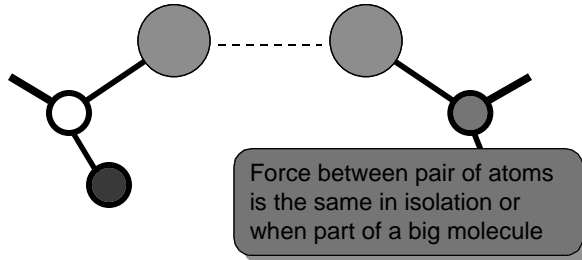


- Alanine
- Proline
- Threonine
- Tryptophan
- Isoleucine

NMR, X-ray and electron crystallography solve structures slowly (1/2-3 yrs.)

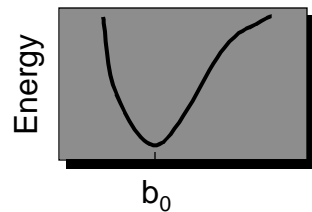
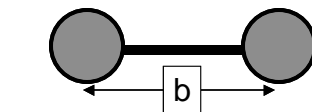
## The "Beads" are Chemically Complex Structures





## ■ Basic assumptions:

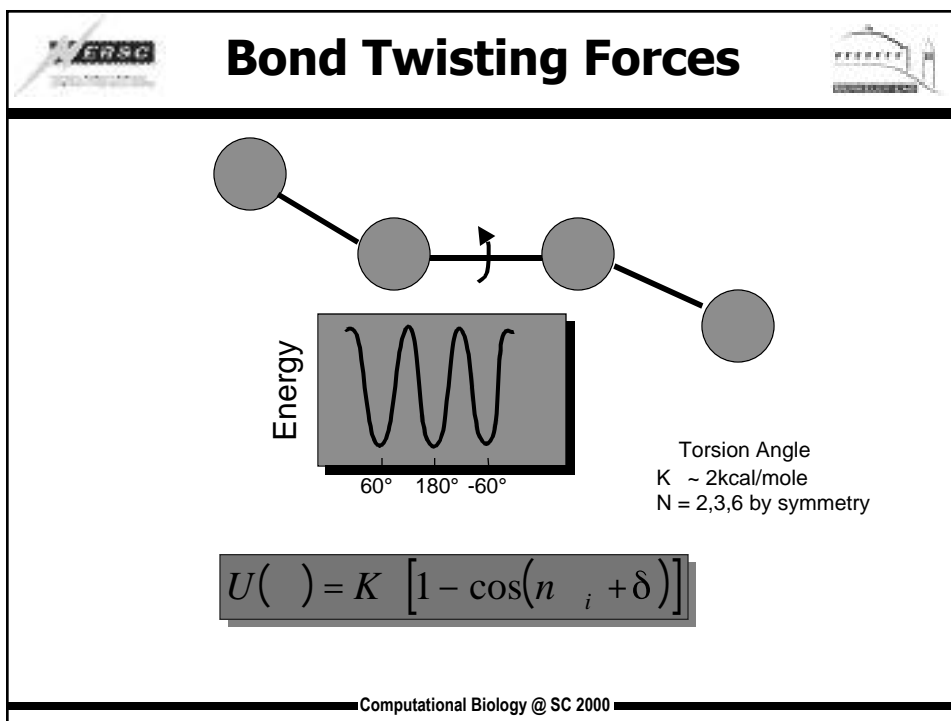
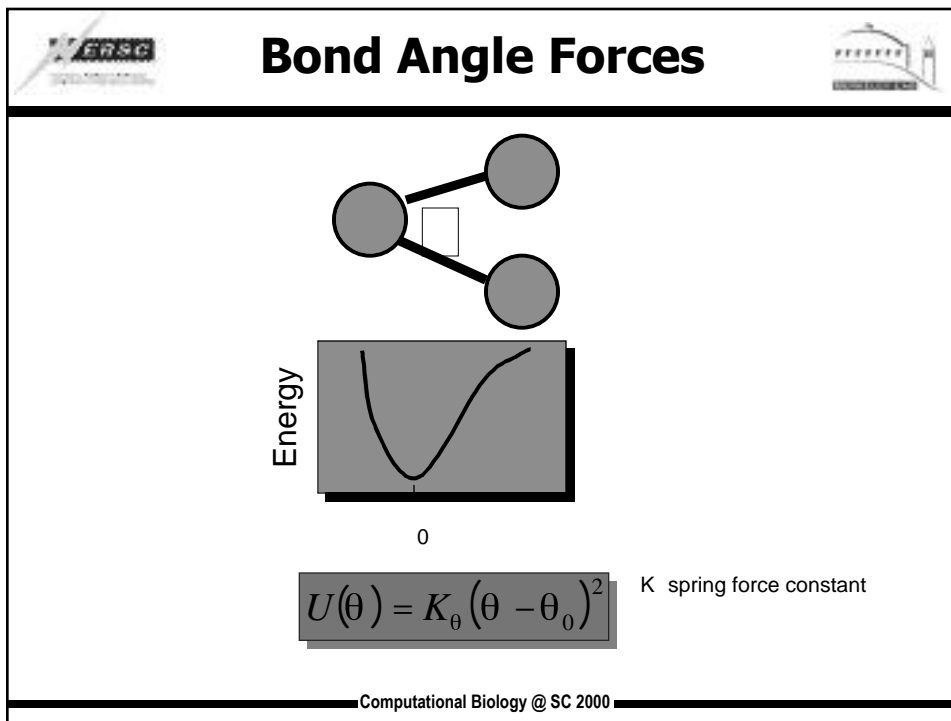
- ✓ Energy contributions are strictly additive
- ✓ Energy is independent of neighbors; transferability
- ✓ Quantum mechanics is insignificant as long as no bonds are broken

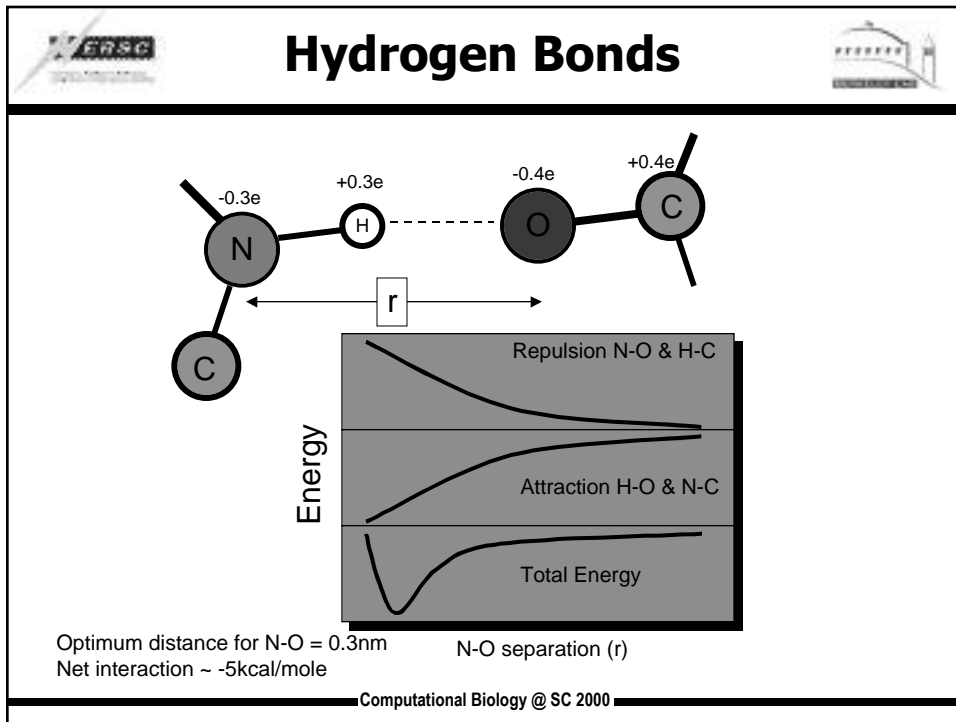



Equilibrium length ~ 0.1-0.2nm

$$U(b) = K_b (b - b_0)^2$$


$K_b$  spring force constant ~ 500kcal/mole Å<sup>2</sup>







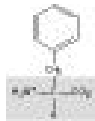


## Scale of Interactions



Interaction	Energy (kcal/mole)
Van der Waals (in water)	-0.1
Hydrogen bond (in water)	-1.0
Torsion barrier (single bond)	~+3.0
Torsion barrier (double bond)	+20.0
Bond breakage	+100.0
Change bond angle by 10°	+2.0
Stretch bond length by 10pm (0.1Å)	+2.5
Thermal energy 300K	0.6

Computational Biology @ SC 2000

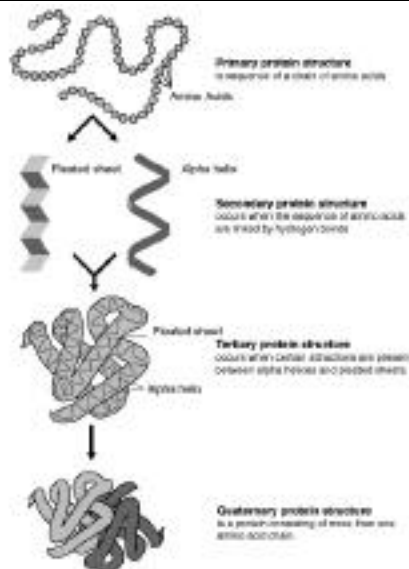
# Aromatic Amino Acids

Amino Acid	pK <sub>a</sub> 's <sup>2</sup>	Pro Structure <sup>1</sup>	Chemical Structure <sup>1</sup>	3-D Structure <sup>3</sup>
Phenylalanine, Phe, F No charge absorbs UV hydrophobic (2.5) Molec. Wt. = 147 Mole % = 3.5	N=9.13 C=1.83 pI=5.48	a =1.16 β =1.33 t =0.59		
Tyrosine, Tyr, Y weak charge absorbs UV hydrogen bonding not hydrophilic (0.08) Molec. Wt. = 163 Mole % = 3.5	N=9.11 C=2.20 R=10.07 pI=5.66	a =0.74 β =1.45 t =0.76		
Tryptophan, Trp, W largest amino acid rarest amino acid no charge absorbs UV hydrogen bonding hydrophobic (1.5) Molec. Wt. = 186 Mole % = 1.1	N=9.39 C=2.38 pI=5.89	a =1.02 β =1.35 t =0.65		

Copyright© Charles S. Gasser 1996

Computational Biology @ SC 2000

# Protein Structure



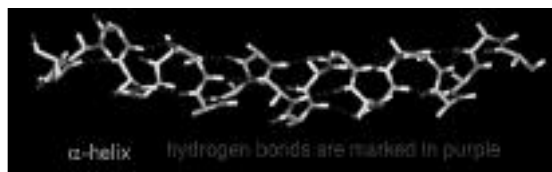
Computational Biology @ SC 2000

- Alpha-helix
- Beta-sheet
- Coil




## ■ Alpha-helix


- ✓ Right-handed alpha helix
- ✓ 3.6 amino acids per turn
- ✓ Most abundant (35%)

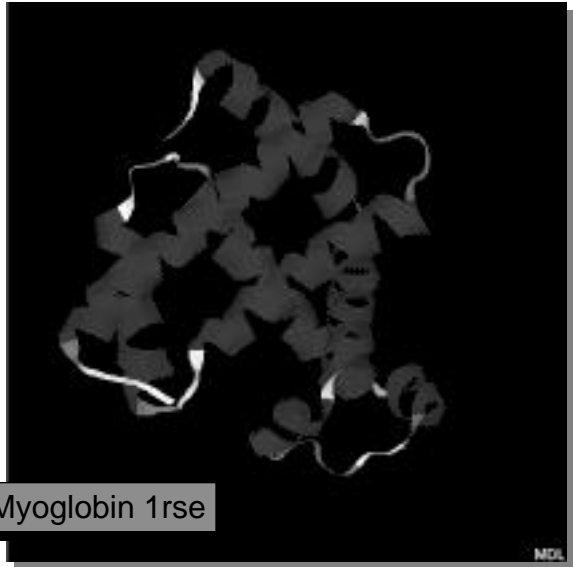






# Alpha Helix







Human Myoglobin 1rse

MOL

Computational Biology @ SC 2000




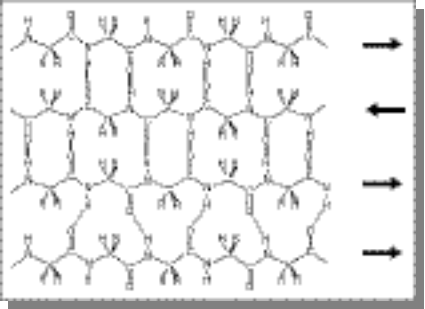
# Beta-Sheet



■ **Beta-sheet**

- ✓ Parallel - antiparallel
- ✓ 25% of proteins

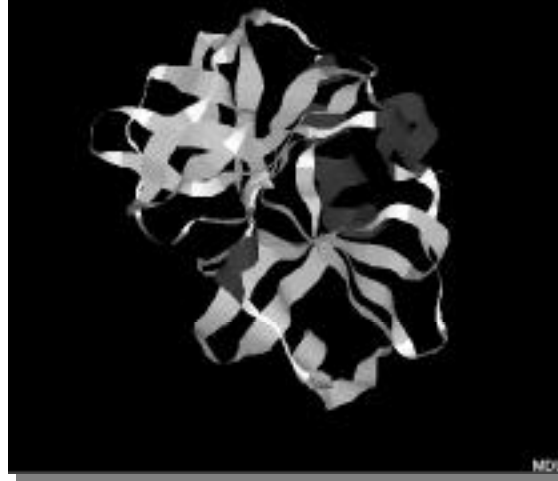




Computational Biology @ SC 2000

## Beta sheets

Human Rhinovirus Protease 3C 1cqq



Computational Biology @ SC 2000

## SCOP: Structural Classification of Proteins

- 1. All alpha proteins (a)
- 2. All beta proteins (b)
- 3. Alpha and beta proteins (a/b)
  - ✓ Mainly parallel beta sheets (beta-alpha-beta units)
- 4. Alpha and beta proteins (a+b)
  - ✓ Mainly antiparallel beta sheets (segregated alpha and beta regions)
- 5. Multi-domain proteins (alpha and beta)
  - ✓ Folds consisting of two or more domains belonging to different classes
- 6. Membrane and cell surface proteins and peptides
  - ✓ Does not include proteins in the immune system
- 7. Small proteins
  - ✓ Usually dominated by metal ligand, heme, and/or disulfide bridges
- 8. Coiled coil proteins
- 9. Low resolution protein structures
- 10. Peptides
- 11. Designed proteins

Computational Biology @ SC 2000

SCOP Classifications			
Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	128	197	296
All beta proteins	87	158	251
Alpha and beta proteins (a/b)	93	153	323
Alpha and beta proteins (a+b)	168	237	345
Multi-domain proteins	25	25	32
Membrane and cell surface proteins	11	17	19
Small proteins	52	72	102
Total	564	859	1368

SCOP: Structural Classification of Proteins. 1.53 release  
 11410 PDB Entries (1 Jul 2000).  
 26219 Domains.  
 Copyright © 1994-2000 The scop authors / scop@mrc-lmb.cam.ac.uk  
 September 2000

Computational Biology @ SC 2000

Protein Fold Recognition, Structure Prediction, and Folding	
<ul style="list-style-type: none"> <li>■ Drawing analogies with known protein structures               <ul style="list-style-type: none"> <li>✓ Sequence homology, Structural Homology</li> <li>✓ Inverse Folding, Threading</li> </ul> </li> <li>■ Ab initio folding: the ability to follow kinetics, mechanism               <ul style="list-style-type: none"> <li>✓ robust objective function</li> <li>✓ severe time-scale problem</li> <li>✓ proper treatment of long-ranged interactions</li> </ul> </li> <li>■ Ab initio prediction: the ability to extrapolate to unknown folds               <ul style="list-style-type: none"> <li>✓ multiple minima problem</li> <li>✓ robust objective function</li> <li>✓ Stochastic Perturbation and Soft Constraints</li> </ul> </li> <li>■ Simplified Models that Capture the Essence of Real Proteins               <ul style="list-style-type: none"> <li>✓ Lattice and Off-Lattice Simulations</li> <li>✓ Off-Lattice Model that Connect to Experiments: Whole Genomes?</li> </ul> </li> </ul>	

Computational Biology @ SC 2000



## Protein Fold Predictions: Neural Network Structure Classifications



- Protein fold predictor based on global descriptors of amino acid sequence
- Empirical prediction using a database of known folds in machine learning
- Databases
  - 3D-ALI (83 folds)
  - SCOP (used ~120 folds)
- Representation of protein sequence in terms of physical, chemical, and structural properties of amino acids
- Feed forward neural network for machine learning

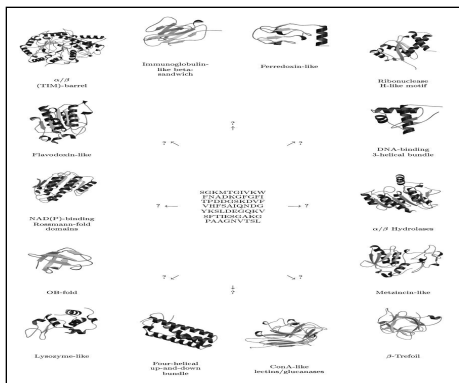
Computational Biology @ SC 2000



## Protein Fold Recognition: Threading




*Sequence Assignments to Protein Fold Topology*  
(David Eisenberg, UCLA)




Take a sequence with unknown structure and align onto structural template of a given fold  
Score how compatible that sequence is based on empirical knowledge of protein structure  
Right now 25-30% of new sequences can be assigned with high confidence to fold class  
100,000's of sequences and 10,000's of structures (each of order  $10^2$ - $10^3$  amino acids long)

Computational Biology @ SC 2000



## Protein Fold Recognition: Threading



---

**Computational Approach:**

*Dynamic programming:* capable of finding optimal alignments if  
 optimal alignments of subsequences can be extended to optimal alignments of whole  
 objective functions that are one-dimensional  $E = V_i + V_{\text{gap}}$

*Complexity:* all to all comparison of sequence to structure scales as  $L^2$   
 Whole human genome:  $10^{13}$  flops

**Improve Objective function:**

*Take into account structural environment*

3D ~~2D~~: dynamic programming,  $L^2$

*Build pairwise or multi-body objective function*


NP-hard if: variable-length gaps and model nonlocal effects such as distance  
 dependence

Recursive dynamic programming, Hidden markov models, stochastic grammars


*Complexity:* all to all comparison of sequence to structure scales as  $L^3$   
 Whole human genome:  $\sim 10^{16}$  flops

---

Computational Biology @ SC 2000




## Computational Protein Folding



---

*One microsecond simulation of a fragment of the protein, Villin. (Duan & Kollman, Science 1998)*



✓

robust objective function

all atom simulation with molecular water present: some structure present

✓

severe time-scale problem

required  $10^9$  energy and force evaluations: parallelization (spatial decomposition)

✗

proper treatment of long-ranged interactions

cut-off interactions at  $8\text{\AA}$ , poor by known simulation standards


✗

Statistics (1 trajectory is anecdotal)


Many trajectories required to characterize kinetics and thermodynamics

---

Computational Biology @ SC 2000



# Computational Protein Folding



---

(1) Size-scaling bottlenecks: Depends on complexity of energy function, V

Empirical (less accurate):  $cN^2$ ; ab initio (more accurate):  $cN^3$  or worse ;  $c \ll C$

empirical force field used

“long-ranged interactions” truncated so  $cM^2$  scaling;  $M < N$

spatial decomposition, linked lists

(2) Time-Scale of motions bottlenecks ( $\Delta t$ )

$$r_i(t + \Delta t) = 2r_i(t) - r_i(t - \Delta t) + \frac{f_i(t)}{m_i} \frac{(\Delta t)^2}{2!} + O[(\Delta t)^4]; v_i(t) = \frac{r_i(t + \Delta t) - r_i(t - \Delta t)}{2 \Delta t} + O[(\Delta t)^3]$$

$$f_i = m_i a_i = - \nabla_i V(r_1, r_2, \dots, r_N)$$

Use timestep commensurate with fastest timescale in your system


bond vibrations: 0.01 Å amplitude:  $10^{-15}$  seconds (1fs)

Shake/Rattle bonds (2fs)


Multiple timescale algorithms (~5fs) (not used here)

---

Computational Biology @ SC 2000



# Ab Initio Protein Structure Prediction



---

Primary Sequence and an Energy function → Tertiary structure

Empirical energy functions:

(1) Detailed, Atomic description: leads to enormous difficulties!

$$V_{MM} = \sum_i^{\text{\# Bonds}} k_b (b_i - b_o)^2 + \sum_i^{\text{\# Angles}} k_\theta (\theta_i - \theta_o)^2 + \sum_i^{\text{\# Improvers}} k_\tau (\tau_i - \tau_o)^2 +$$

$$\sum_i^{\text{\# dihedrals}} k_\phi [1 + \cos(n\phi + \delta)] + \sum_{i < j}^{\text{\# atoms} \text{\# atoms}} \frac{q_i q_j}{r_{ij}} + \epsilon_{ij} \frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{ij}^6}{r_{ij}^6} + \sum_i^{\text{\# atoms}} \sigma A$$

(1) Multiple minima problem is fierce


Find a way to effectively overcome the multiple minima problem

(2) Objective Functions: Replaceable algorithmic component?


Global energy minimum should be native structure, misfolds higher in energy

---

Computational Biology @ SC 2000





# The Objective (Energy) Function



---

**Empirical Protein Force Fields: AMBER, CHARMM, ECEPP**  
“gas phase”

CATH protein classification: <http://pdb.pdb.bnl.gov/bsm/cath>

$\alpha$ -helical sequence/  $\beta$ -sheet structure       $\beta$ -sheet sequence/ $\alpha$ -helical structure

**Energies the same! Makes energy minimization difficult!**


**Add penalty for exposing hydrophobic surface: favors more compact structures**

$E_{\text{native folds}} < E_{\text{misfolds}}$  for a few test cases


**Solvent accessible surface area functions: Numerically difficult to use in optimization**

---

Computational Biology @ SC 2000



# Neural Networks for 2° Structure Prediction

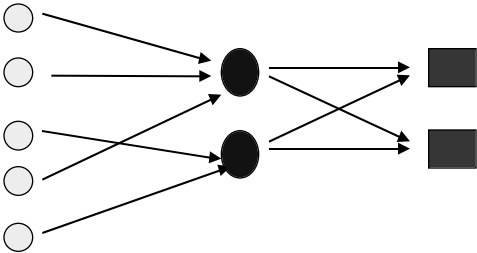


---

○ Input units represent amino acid sequence

● Hidden units map sequence to structure

■ Output Units represent secondary structure class (helix, sheet, coil)



→ Weights are optimizable variables that are trained on database of proteins



Poorly designed networks result in overfitting, inadequate generalization to test set

Neural network design

- input and output representation
- number of hidden neurons
- weight connection patterns that detect structural features

---

Computational Biology @ SC 2000

## Neural Network Results

No sequence homology through multiple alignments

Train	Test
Total predicted correctly = 66%	Total predicted correctly = 62.5%
Helix: 51% $C_a=0.42$	Helix: 48% $C_a=0.38$
Sheet: 38% $C_b=0.39$	Sheet: 28% $C_b=0.31$
Coil: 82% $C_c=0.36$	Coil: 84% $C_c=0.35$



Network with Design: Yu and Head-Gordon, Phys. Rev. E 1995

Train	Test
Total predicted correctly = 67%	Total predicted correctly = 66.5%
Helix: 66% $C_a=0.52$	Helix: 64% $C_a=0.48$
Sheet: 63% $C_b=0.46$	Sheet: 53% $C_b=0.43$
Coil: 69% $C_c=0.43$	Coil: 73% $C_c=0.44$

Combine networks of Yu and Head-Gordon with multiple alignments

---

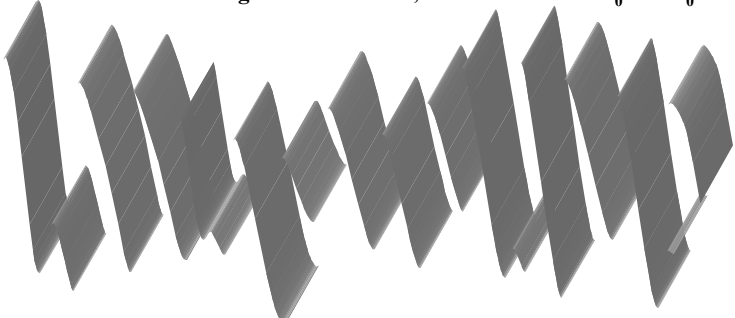
Computational Biology @ SC 2000

## Neural Networks Used To Guide Global Optimization Methods

Generate expanded tree of configurations

Predicted coil residues: generate random, dissimilar sets of  $\phi_0$  and  $\psi_0$



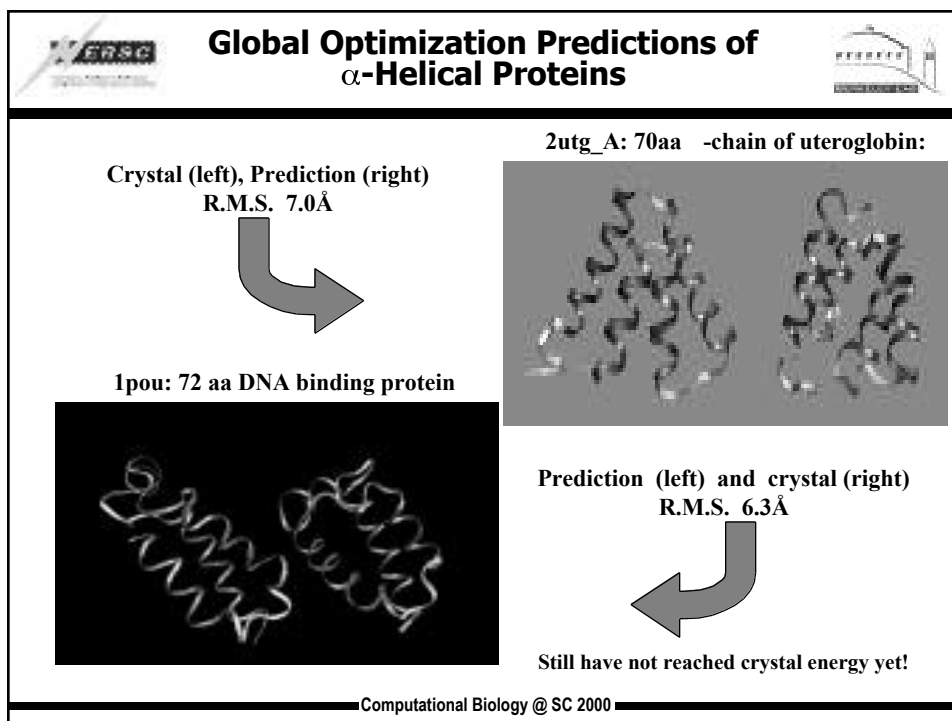
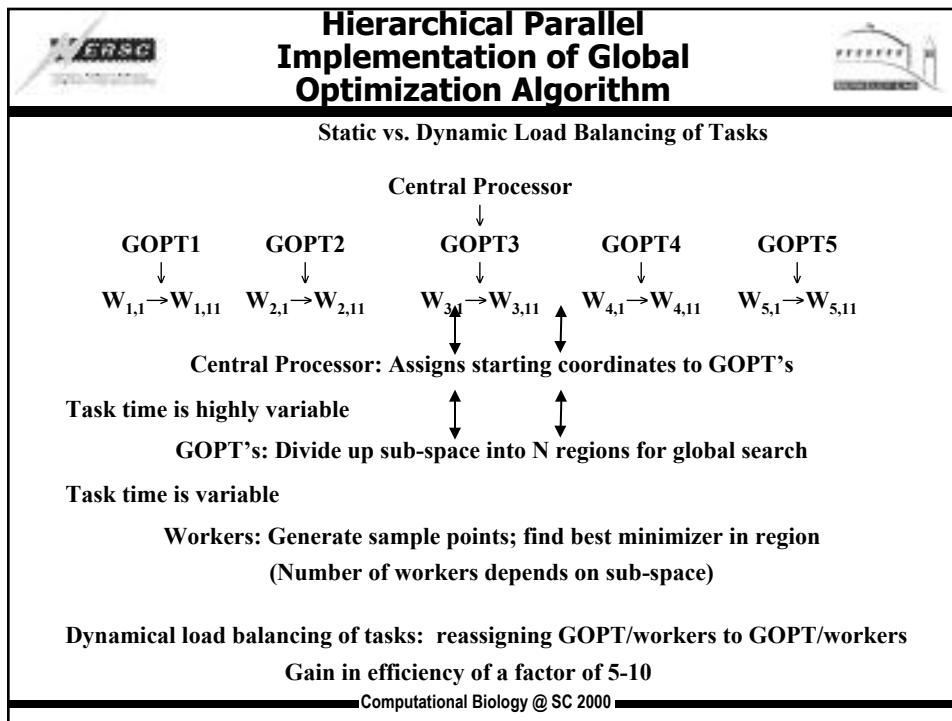
Explore tree configuration in depth:


Global Optimization in sub-space of coil residues: walk through barriers, move downhill

---


Computational Biology @ SC 2000

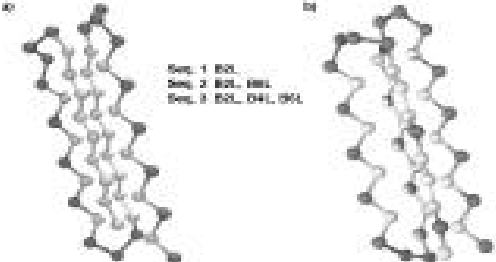






## Simplified Models for Simulating Protein Folding






(A) Ser. 1 R2L  
 (B) Ser. 2 R2L, L84L  
 (C) Ser. 3 R2L, D41L, Q64L


Simplifies the “real” energy surface topology sufficiently that you can do

- (1) Statistics ✓  
Can do many trajectories to converge kinetics and thermodynamics
- (2) severe time-scale problem ✓  
characterize full folding pathway: mechanism, kinetics, thermodynamics
- (3) proper treatment of long-ranged interactions ✓  
all interactions are evaluated; no explicit electrostatics
- (4) robust objective function?  
good comparison to experiments

Computational Biology @ SC 2000



## Acknowledgements



Teresa Head-Gordon, Physical Biosciences Division, LBNL

Silvia Crivelli, Physical Biosciences and NERSC Divisions, LBNL

Betty Eskow, Richard Byrd, Bobby Schnabel, Dept. Computer Science,  
U. Colorado

Jon M. Sorenson, NSF Graduate Fellow, Dept. Chemistry UCB

Greg Hura, Graduate Group in Biophysics, UCB

Alan K. Soper, Rutherford Appleton Laboratory, UK

Alexander Pertsemlidis, Dept. of Biochemistry, U. Texas Southwestern Medical  
Center

Robert M. Glaeser, Mol. & Cell Biology, UCB and Life Sciences Division, LBNL

Computational Biology @ SC 2000



# Structure-Based Drug Discovery


Brian K. Shoichet, Ph.D  
Northwestern University, Dept of MPBC  
303 E. Chicago Ave, Chicago, IL 60611-3008  
Nov 15, 1999




## Problems in Structure-Based Inhibitor Discovery & Design

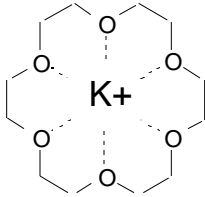


- **Balance of forces in binding**
  - **Energies in condensed phases**
    - ✓ interaction energies
    - ✓ desolvation
- **Problem scales badly with degrees of freedom**
  - **Configuration**
    - ✓ configs = (prot-features)<sup>4</sup> X (lig-features)<sup>4</sup>
  - **Conformation**
    - ✓ Ligand & Protein, confs = 3lbonds X 3pbonds
- **Sampling chemical space (scales very badly)**
- **Defining binding sites**

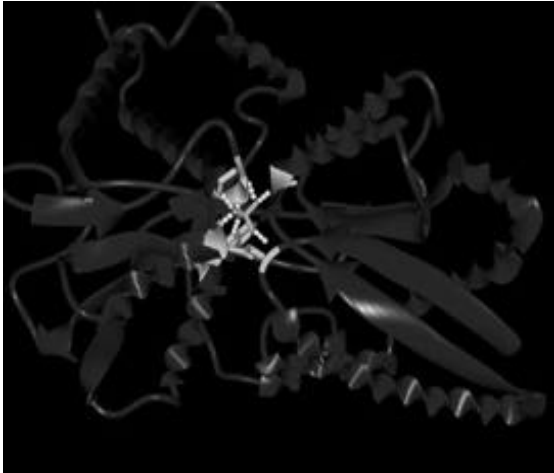


## The Pros & Cons of Proteins






18 - Crown-6




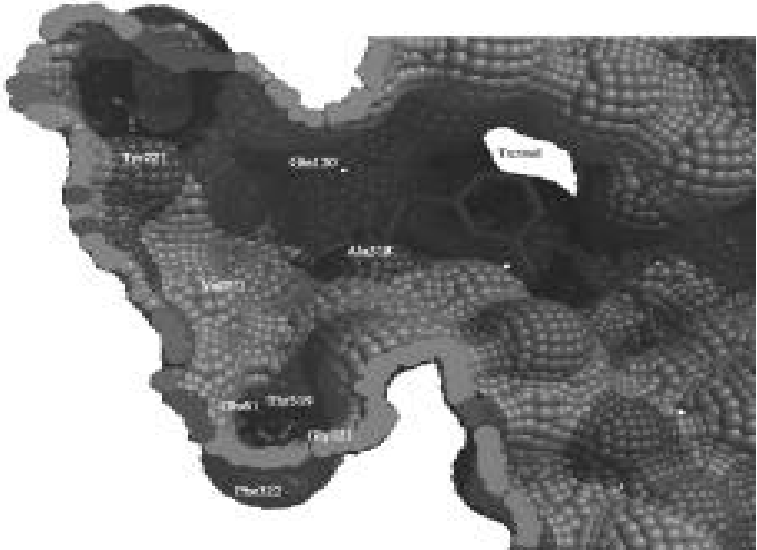
sulfate binding protein

Computational Biology @ SC 2000



## Conserved Residues, Ordered Structure, Function Unknown





Computational Biology @ SC 2000

# Inhibitor Discovery or Design?

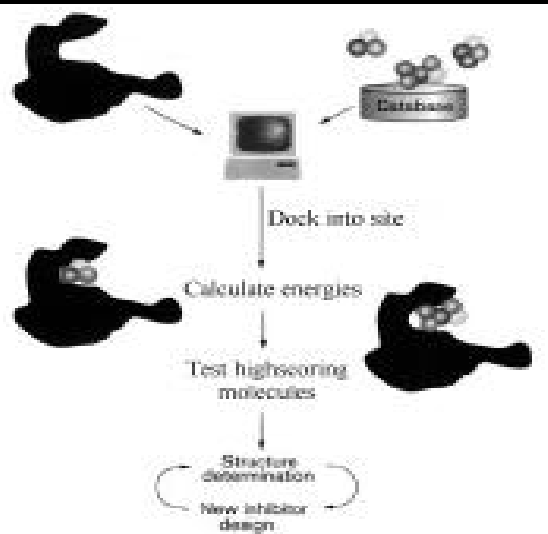
## ■ Design ligands

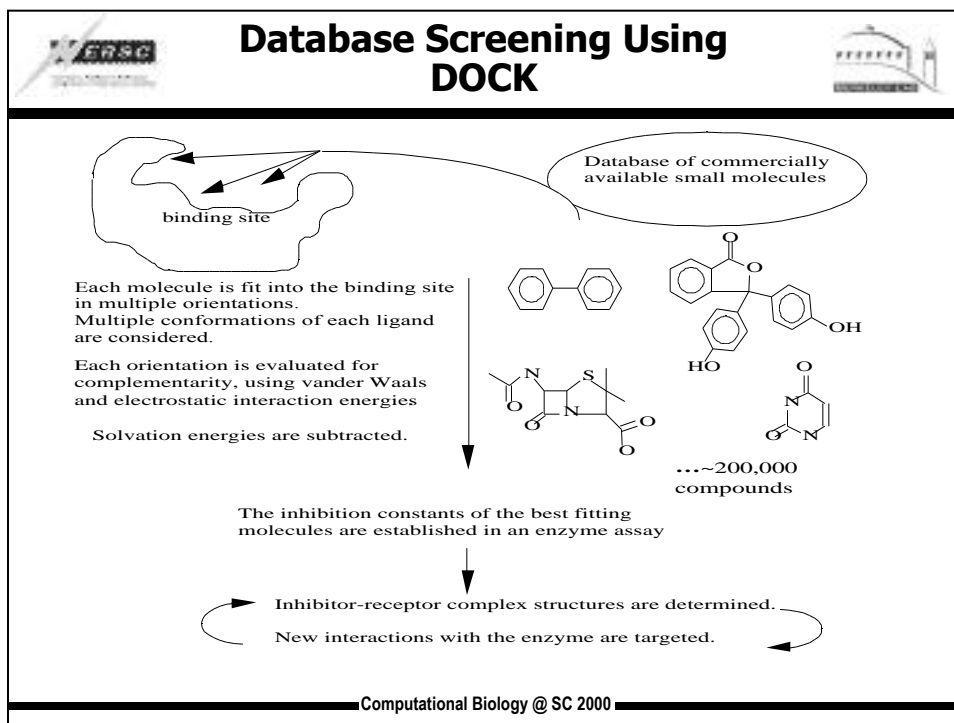
- Ludi (Bohm)
- Grow (Moon & Howe)
- Builder (Roe & Kuntz)
- MCSS-Hook (Miranker & Karplus)
- SMOG (DeWitte & Shakhovitch)
- Others...


## ■ Discover Ligands

- DOCK (Kuntz, et al., Shoichet)
- CAVEAT (Bartlett)
- Monte Carlo (Hart & Read)
- AutoDock (Goodsell & Olson)
- SPECITOPE (Kuhn et al)
- Others...


# Screening Databases by Molecular Docking

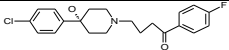
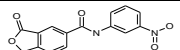
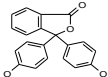
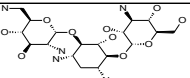


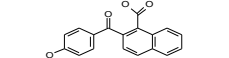

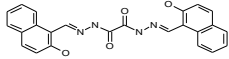
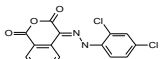





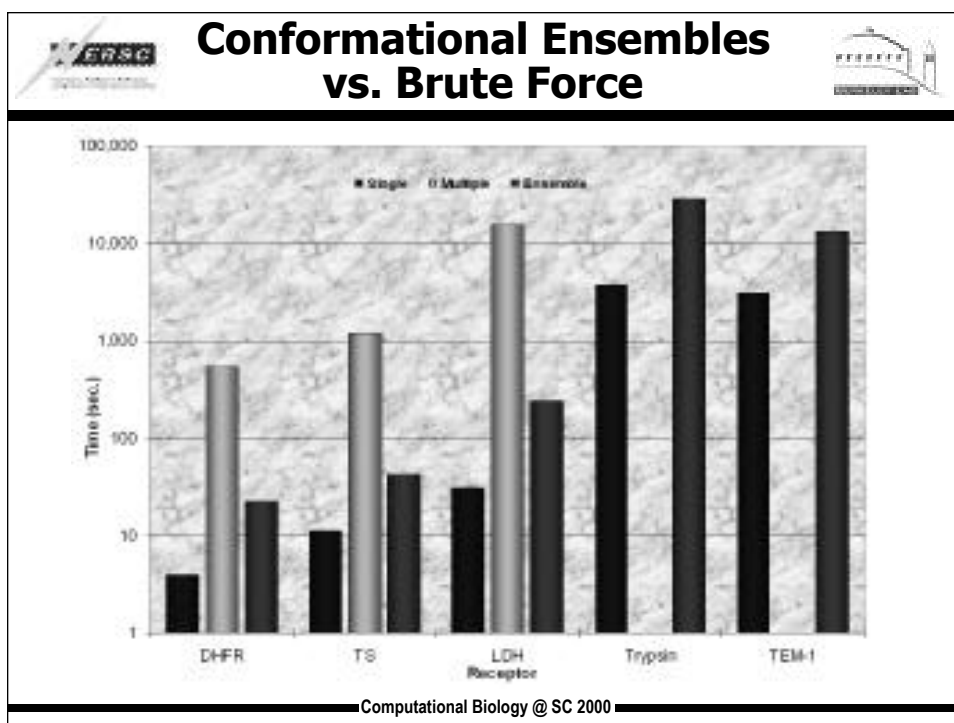
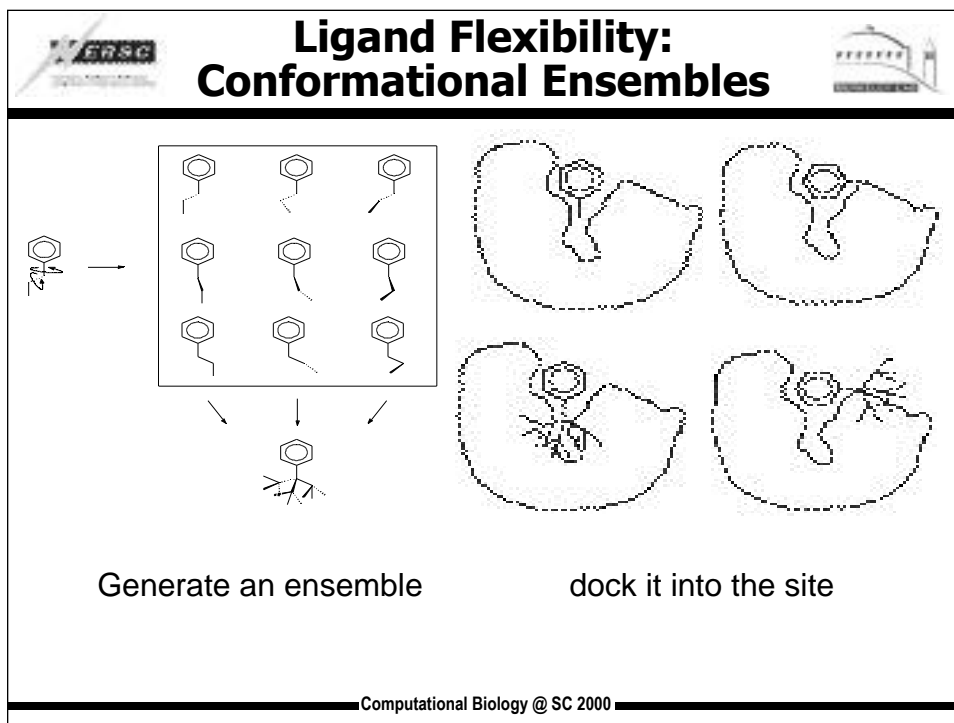


## Novel Ligand Discovery Using Molecular Docking

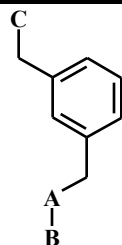


Receptor	Lead from molecular docking	Receptor	Lead from molecular docking
HIV protease		HGXPRtase	
thymidylate synthase			
hemagglutinin		-lactamase	
cercarial elastase		-lactamase	
malarial protease		thymidylate synthase	
CD4-gp120	unpublished		

Computational Biology @ SC 2000

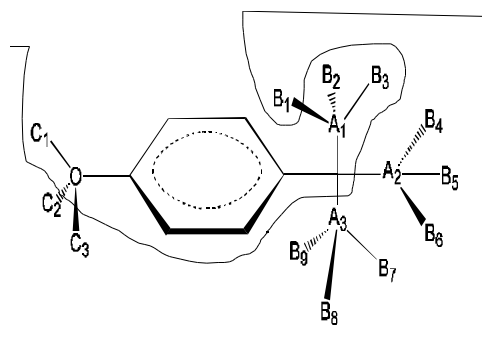


# Hierarchical Docking



Flexible docking:  
27 confs  
x3 atoms  
81 atom positions

Hierarchical docking:  
27 confs  
 $3C + 3A + 9B$   
15 atom positions







# Computational Phylogenetics

Craig Stewart  
stewart@iu.edu  
Indiana University



## Outline



- Evolution & Phylogenetics
- Why is this an HPC problem?
- Alignment (brief)
- Summary of methods and software for phylogenetics
- One example in detail: Maximum Likelihood analysis with fastDNAmI
- Some interesting results and challenges for the future
- Caveat: this is an introduction, not an exhaustive review.



# Phylogeny

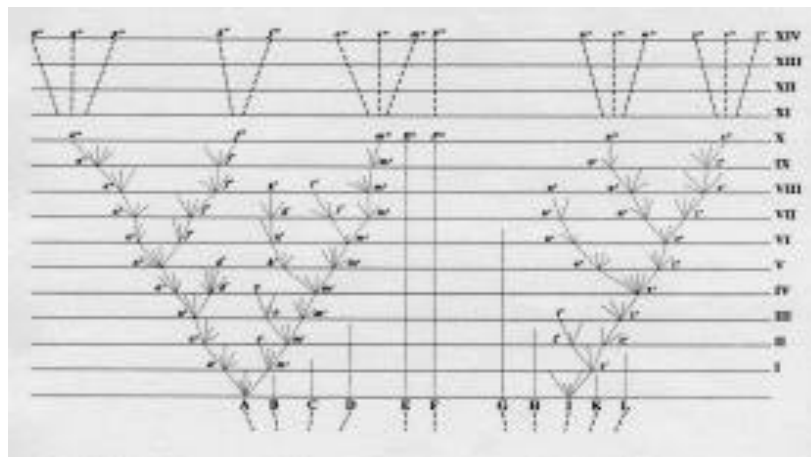


- Evolution is an explicitly historical branch of biology, one in which the subjects are active players in the historical changes.
- A phylogeny, or phylogenetic tree, is a way of depicting evolutionary relationships among organisms, genes, or gene products.
- Modern evolutionary theory began with Darwin's Origin of Species, which included one figure – an evolutionary tree

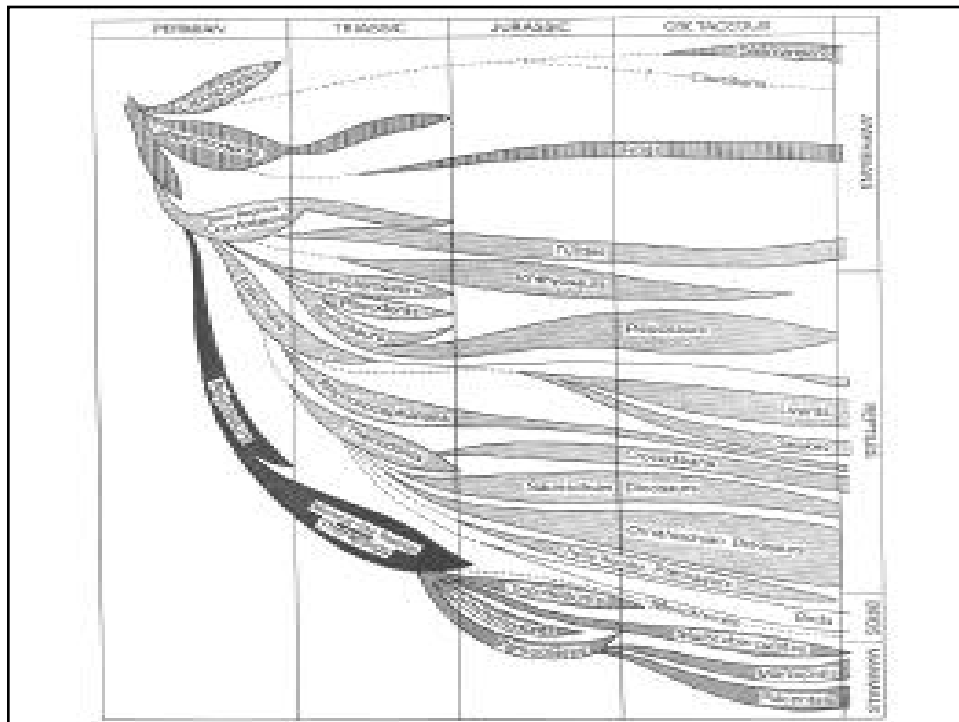
Computational Biology @ SC 2000




## Origin of Species, Figure 1

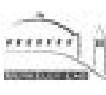


Computational Biology @ SC 2000

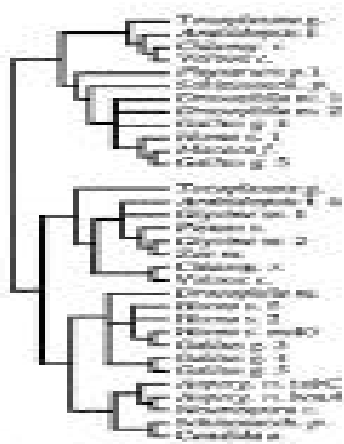




## Building Phylogenetic Trees



- Goal: an objective means by which phylogenetic trees can be estimated in tolerable amounts of wall-clock time, producing phylogenetic trees with measures of their uncertainty



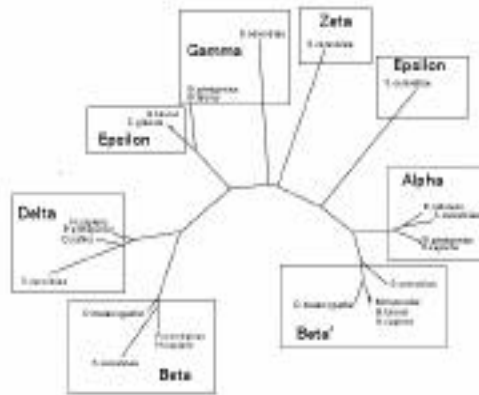
Computational Biology @ SC 2000



# Basic Evolutionary Biology



- All evolutionary changes are described as bifurcating trees
  - evolutionary relationships among genes or gene products (trees of paralogues)
  - evolutionary relationships among organisms (trees of orthologues)



Computational Biology @ SC 2000



## Why?



- **Curiosity:** Anyone who as a child wandered through the dinosaur section of a natural history museum understands the inherent intellectual attraction of evolutionary biology
- **Theoretical uses:** testing hypotheses in evolutionary biology
- **Practical uses:**
  - Medicine
  - Environmental management (biodiversity maintenance)

Computational Biology @ SC 2000



## Reconstructing history from DNA sequences

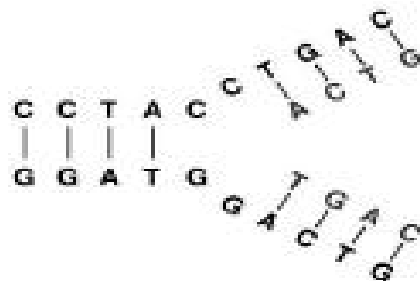


- DNA changes over time; much of this change is not expressed
- Changes in unexpressed DNA can be modeled as Markov processes
- By comparing similar regions of DNA from different organisms (or different genes) one can infer the phylogenetic tree and evolutionary history that seems the best explanation of the current situation

Computational Biology @ SC 2000



## DNA replication



**Purines:**

**Pyrimidines:**

**Adenine & Guanine**

**Thymine & Cytosine**

Computational Biology @ SC 2000



## Changes in genetic information over time



- **Point mutations**

**DNA – sequences of the 4 nucleotides**

CCTCTGAC

vs

TCTCCGAC

**Protein – sequences of the 20 amino acids**

GSAQVKGHGKK

vs

GNPKVKAHGKK

- **Insertions and deletions**

**DNA**

CCTCT+GAC

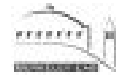
vs

CCTCTTGAC

Computational Biology @ SC 2000



## Sequences available

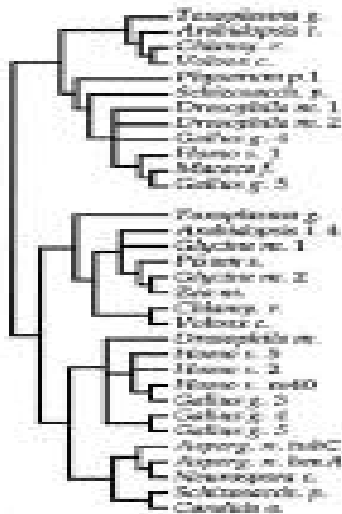


- **DNA (sequences are series of the base molecules; aligned sequences will also contain +s for gaps)**
- **Amino acid sequences (series of letters indicating the 20 amino acids). Computational challenges more severe than with DNA sequences.**
- **RNA**
- **The availability of data at present exceeds the ability of researchers to analyze it!**

Computational Biology @ SC 2000



## Why is tree-building a HPC problem?



- The number of bifurcating unrooted trees for  $n$  taxa is  $(2n-5)! / (n-3)! 2^{n-3}$
- for 50 taxa the number of possible trees is  $\sim 1074$ ; most scientists are interested in much larger problems
- The number of rooted trees is  $(2n-5)!$

Computational Biology @ SC 2000



## Alignment



- To build trees one compares and relates 'similar' segments of genetic data. Getting 'similar' right is absolutely critical!
- Methods:
  - dynamic programming
  - Hidden Markov Models
  - Pattern matching
- Some alignment packages:
  - BLAST  
<http://www.ncbi.nlm.nih.gov/BLAST/>
  - FASTA  
<http://gcg.nhri.org.tw/fasta.html>
  - MUSCA  
<http://www.research.ibm.com/bioinformatics/home>

Computational Biology @ SC 2000



## Matching cost function



GCTAAATTC

++ x x

GC AAGTT

- Penalize for mismatches, for opening of gap, and for gap length
- This approach assumes independence of loci: good assumption for DNA, some problems with respect to amino acids, significant problems with RNA

Computational Biology @ SC 2000



## Example of aligned sequences



Thermotoga	ATTGCCCCA	GAAATTAAAG	CAAAAACCCC	AGTAAGTTGG	GGATGGCAA
Tthermophi	ATTGCCCCA	GGGGTTCCCG	CAAAAACCCC	AGTAAGTTGG	GGATGGCAGG
Taquaticus	ATTGCCCCA	GGGGTTCCCG	CAAAAACCCC	AGTAAGTTGG	GGATGGCAGG G
deinon	ATTGCCCCA	GGGATTCCCG	CAAAAACCCC	AGTAAGTTGG	GGATGGCAGG G
Chlamydi	ATTTTCCCCA	GAAATTCCCG	AAAAAACCCC	AATAAATTGG	GGATGGCAGG
flexistips	ATTTTCCCCA	CAAAAAAAG	AAAAAACCCC	AGTAAGTTGG	GGATGGCAGG
borrelia-b	ATTGCCCCA	GAAATTAAAG	CAAAAACCCC	AATAAATTGG	GGATGGCAGG
bacterioide	ATTGCCCCA	GAAATTCCCG	CAAAAACCCC	AGTAAATTGG	GGATGGCAGG GG
Pseudom	ATTGCCCCA	GGGATTCCCG	CAAAAACCCC	AGTAAGTTGG	GGATGGCAGG G
ecoli----	GTTTCCCCA	GAAATTCCCG	CAAAAACCCC	AGTAAGTTGG	GGATGGCAGG
salmonella	+++++				
shewanella	GTTTCCCCA	GCCATTCCCG	TAAAAACCCC	AGTAAGTTGG	GGATGGCAGG
bacillus--	ATTGCCCCA	GAAATTCCCG	CAAAAACCCC	AGCAAATTGG	GGATGGCAGG G
myco-gentl	ATTGCCCCG	GAAATTCCCG	CAAAAACCCC	AGTAAGTTGG	GGATGGCAA

Computational Biology @ SC 2000





## Phylogenetic methodologies



- Define a specific series of steps to produce the 'best' tree
  - Pair-group cluster analyses
  - Fast, but tend not to address underlying evolutionary mechanisms
- Define criteria for comparing different trees and judging which is better. Two steps:
  - Define the objective function (evolutionary biology)
  - Generate and compare trees (computation)
- All of the techniques described produce an unrooted tree.
- The trees produced likewise describe relationships among extant taxa, not the progress of evolution over time.

Computational Biology @ SC 2000



## Distance-based Tree-building methods



- Aligned sequences are compared, and analysis is based on the differences between sequences, rather than the original sequence data.
- Less computationally intensive than character-based methods
- Tend to be problematic when sequences are highly divergent

Computational Biology @ SC 2000



## Distance-based Tree building methods, 2

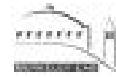


- **Cluster analysis.** Most common variant is Unweighted Pair Group Method with Arithmetic Mean (UPGMA) – join two closest neighbors, average pair, keep going. Problematic when highly diverged sequences are involved
- **Additive tree methods** – built on assumption that the lengths of branches can be summed to create some measure of overall evolution.
  - **Fitch-Margoliash (FM)** – minimizes squared deviation between observed data and inferred tree.
  - **Minimum evolution (ME)** – finds shortest tree consistent with data
- **Of the distance methods, ME is the most widely implemented in computer programs**

Computational Biology @ SC 2000



## Character-based methods



- **Use character data (actual sequences) rather than distance data**
- **Maximum parsimony.** Creates shortest tree – one with fewest changes. Inter-site rate heterogeneity creates difficulties for this approach.
- **Maximum likelihood.** Searches for the evolutionary model that has the highest likelihood value given the data. In simulation studies ML tends to outperform others, but is also computationally intensive.

Computational Biology @ SC 2000



## Rooting trees



- If the assumption of a constant molecular clock holds, then the root is the midpoint of the longest span across the tree.
- Sometimes done by including an 'outgroup' in the analysis
- Remember that the trees produced from sequence data are fundamentally different than a historical evolutionary tree

Computational Biology @ SC 2000



## Evaluating trees



- Once a phylogenetic tree has been produced by some means, how do you test whether or not the tree represents evolutionary change, or just the results of a mathematical technique applied to a set of random data? These methods below can be used to perform a statistical significance test.
- Significance tests for MP trees:
  - Skewness tests. MP tree lengths produced from random data should be symmetric; tree lengths produced from data sets with real signal should be skewed.
- Significance tests for distance, MP, and ML trees:
  - Bootstrap. Recalculate trees using multiple samples from same data with resampling.
  - Jackknife. Recalculate trees using subsampling
- All of these methods are topics of active debate

Computational Biology @ SC 2000



## Phylogenetic software



- **Phylip.** (J. Felsenstein). Collection of software packages that cover most types of analysis. One of the most popular software collections. Free.
- **PAUP.** (D. Swofford). Parsimony, distance, and ML methods. Also one of the most popular software collections. Not free, but not expensive.
- **PAML.** (Ziheng Yang). Maximum likelihood methods for DNA and proteins. Not as well suited for tree searching, but performs several analyses not generally available. Free.
- **fastDNAm.** (G. Olsen). Maximum likelihood method for DNA; becoming one of the more popular ML packages. MPI version available soon; well suited to tree searching in large data sets. Free.

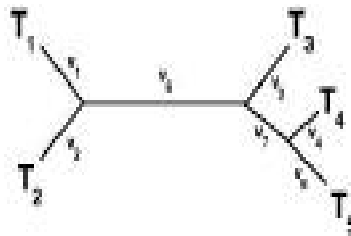
Computational Biology @ SC 2000



## More on Maximum Likelihood methods



- Typical statistical inference: calculate probability of data given the hypothesis
- Tree, branch lengths, and associated likelihood values all calculated from the data.
- Likelihood values used to compare trees and determine which is best



Computational Biology @ SC 2000



## Stochastic change of DNA



- Markov process, independent for each site: 4 x 4 matrix for DNA, 20 x 20 for amino acids

	A	C	G	T
A	$p(A \rightarrow A)$	$p(A \rightarrow C)$	$p(A \rightarrow G)$ ...	
C	$p(C \rightarrow A)$	$p(C \rightarrow C)$	$p(C \rightarrow G)$ ...	
G	.			
T	.			

- Transitions more probable than transversions.
- Must account for heterogeneity in substitution rates among sites (DNArates – Olsen)

Computational Biology @ SC 2000



## fastDNAmI



- Developed by Gary Olsen
- Derived from Felsensteins's PHYLIP programs
- One of the more commonly used ML methods
- The first phylogenetic software implemented in a parallel program (at Argonne National Laboratory, using P4 libraries)
- Olsen, G.J., et al. 1994. fastDNAmI: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Computer Applications in Biosciences 10: 41-48
- MPI version produced in collaboration with Indiana University will be available soon

Computational Biology @ SC 2000



## fastDNAmI algorithm

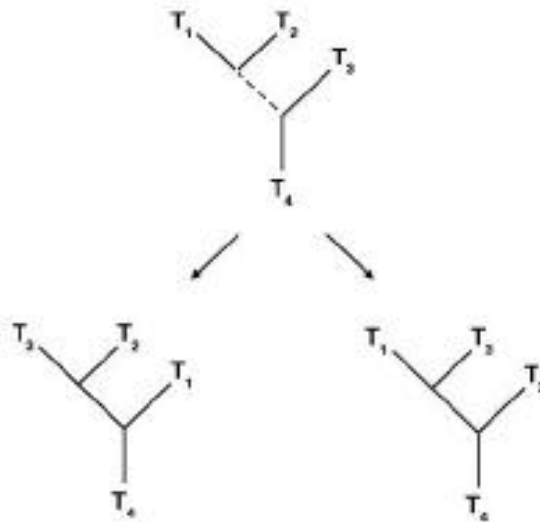


- Compute the optimal tree for three taxa (chosen randomly) - only one topology possible
- Randomly pick another taxon, and consider each of the  $2i-5$  trees possible by adding this taxon into the first, three-taxa tree.
- Keep the best (maximum likelihood tree)
- Local branch rearrangement: move any subtree to a neighboring branch ( $2i-6$  possibilities)
- Keep best resulting tree
- Repeat this step until local swapping no longer improves likelihood value

Computational Biology @ SC 2000



## Local branch rearrangement diagram



Computational Biology @ SC 2000



## fastDNAmI algorithm con't: Iterate

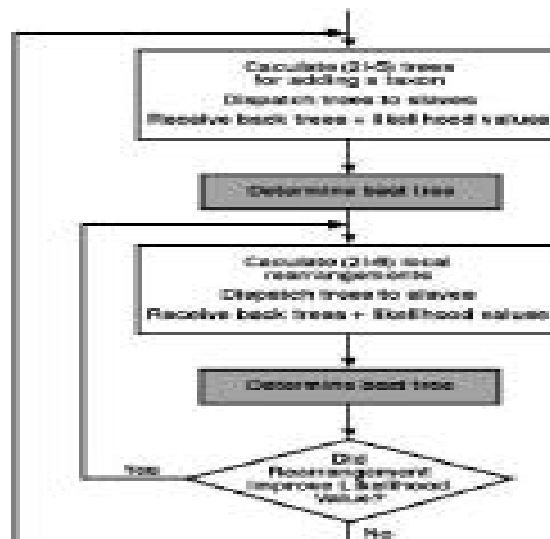


- Get sequence data for next taxon
- Add new taxa (2i-5)
- Keep best
- Local rearrangements (2i-6)
- Keep best
- Keep going....
- When all taxa have been added, perform a full tree check

Computational Biology @ SC 2000



## Overview of parallel program flow



Computational Biology @ SC 2000

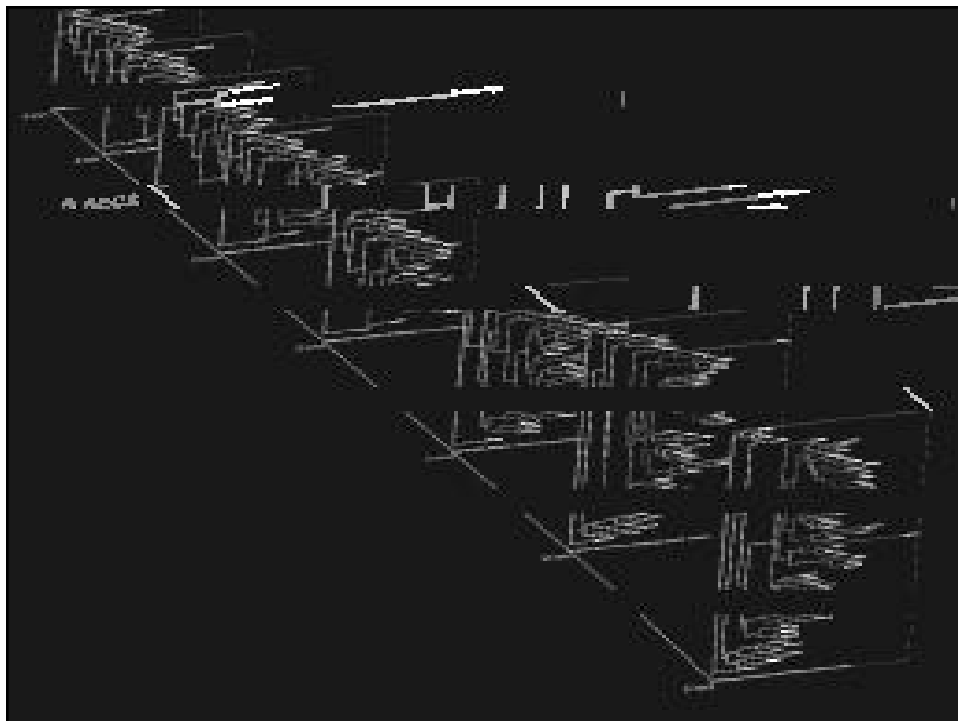


## Because of local effects....

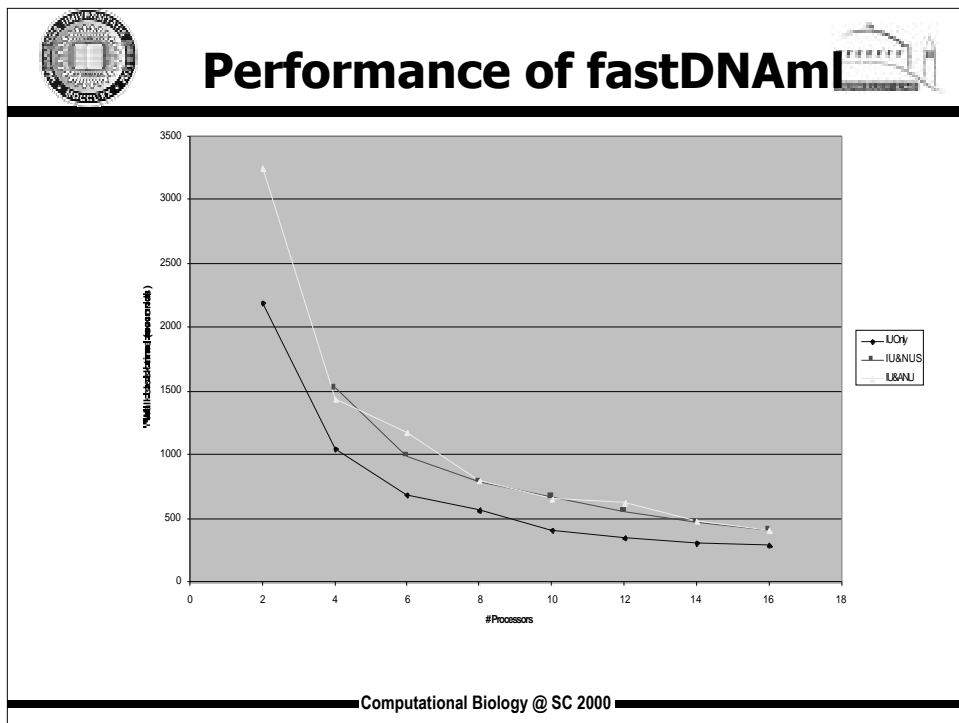
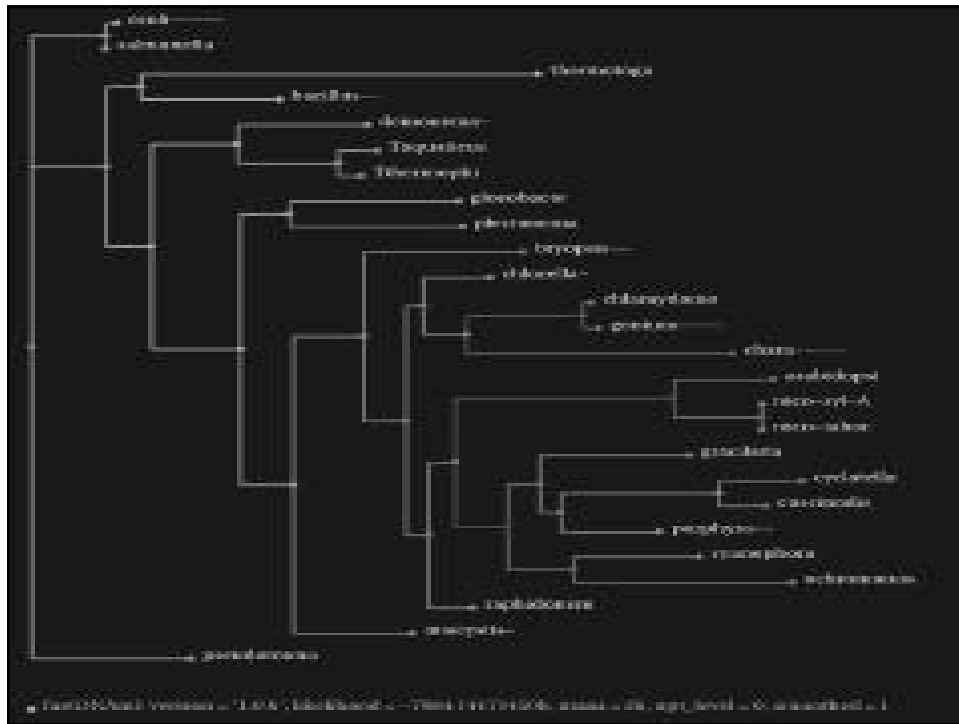


- Where you end up sometimes depends on where you start
- This process searches a huge space of possible trees, and is thus dependent upon the randomly selected initial taxa
- Can get stuck in local optimum, rather than global
- Must do multiple runs with different randomizations of taxon entry order, and compare the results
- Similar trees and likelihood values provide some confidence, but still the space of all possible trees has not been searched extensively

Computational Biology @ SC 2000









## Applications & Interesting examples

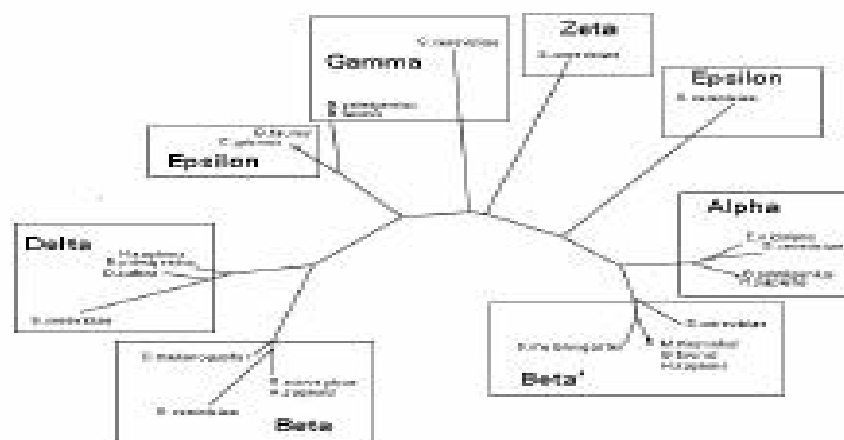


- Better understanding of evolution  
(Ceolocanths, cyanobacterial origin of plastids)
- Maintenance of biodiversity
- Medicine & molecular biology
  - our cousins, the fungi
  - Cytoplasmic coat proteins
  - HIV

Computational Biology @ SC 2000



## Cytoplasmic Coat Proteins



Computational Biology @ SC 2000



# HIV



- Where did HIV come from, and how recent is it?
- Korber, et al. 2000. Timing the ancestor of the HIV-1 pandemic strains. *Science* 288:1789. (Online at [www.sciencemag.org/cgi/content/full/288/5472/1789](http://www.sciencemag.org/cgi/content/full/288/5472/1789))
- Used completed HIV sequences from 159 individuals with known sampling dates (including one from 1959)
- Used a general-reversible (REV) base substitution model, accounting for different site-specific rates of evolution and base frequencies biased in favor of adenosine. Used modified version of fastDNAmI.
- Used SIV as an outgroup
- Last common ancestor of main group of HIV-1 was 1931 (95% confidence interval: 1915-1941). Supports hypothesis that HIV has been around for some time and simply took a while to be common enough to be noticed.

Computational Biology @ SC 2000



# Challenges for future



- HPC implementations of more phylogenetic techniques
- Better treatment of insertions and deletions (indels)
- Algorithms for more thorough searching of treespaces in incremental tree building processes (keep best n trees and keep looking)
- Techniques for not shaking the whole tree (that is, adding a taxa to a tree in a fashion that acknowledges damping of effect as you travel away from altered part of tree)
- Use of high-throughput techniques

Computational Biology @ SC 2000



## Acknowledgements



- The phylogeny depicted in slide 5 is taken from E. Colbert. 1965. The age of reptiles. W.W. Norton, NY, NY.
- Some of the tree diagrams were adapted from Olsen et al. 1994.
- Les Teach [IU] created all other graphics for this talk
- IU's work on parallel versions of fastDNAmI has been facilitated by Shared University Research grants from IBM, Inc.
- IU's work with fastDNAmI would be impossible without our collaboration with Gary Olsen, U. of Illinois, the creator of this program.

Computational Biology @ SC 2000



## References



- Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* 17:368-376
- Baxevanis, A.D., and B.F.F. Ouellette. 1998. *Bioinformatics: a practical guide to the analysis of genes and proteins*. Wiley-Interscience, NY.
- Swofford, D.L., and G.J. Olsen. Phylogeny reconstruction. pp. 411-501 IN D.M. Nillis & C. Mority (eds). *Molecular systematics*. Sinauer Associates, Sunderland, MA.
- Durbin, R. et al. 1998. *Biological sequence analysis*. Cambridge University Press, Cambridge, UK.
- [www.ucmp.berkeley.edu/subway/phylogen](http://www.ucmp.berkeley.edu/subway/phylogen)
- [evolution.genetics.washington.edu/phylip/software](http://evolution.genetics.washington.edu/phylip/software)
- <http://www.indiana.edu/uits/~rac>

Computational Biology @ SC 2000



## URLs for phylogenetic software



- **Phylip**  
[evolution.genetics.washington.edu/phylip/software.html](http://evolution.genetics.washington.edu/phylip/software.html)
- **PAUP**  
[www.lms.si.edu/PAUP/index.html](http://www.lms.si.edu/PAUP/index.html)
- **PAML**  
[abacus.gene.ucl.ac.uk/software/paml.html](http://abacus.gene.ucl.ac.uk/software/paml.html)
- **fastDNAmI**  
[geta.life.uiuc.edu/~gary/](http://geta.life.uiuc.edu/~gary/)

Computational Biology @ SC 2000



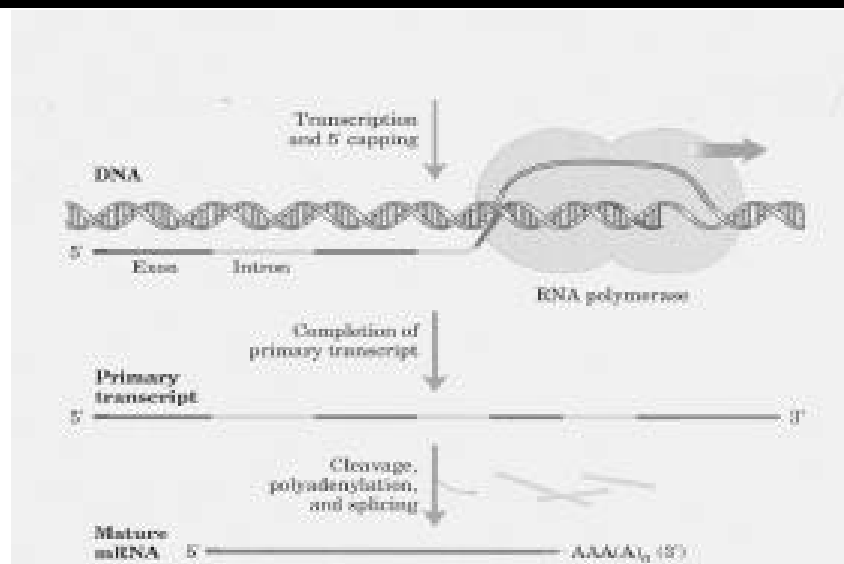
## Specialized biological databases and their role in building models of regulation


Inna Dubchak  
[ILDubchak@lbl.gov](mailto:ILDubchak@lbl.gov)  
NERSC

# Overview of alternative splicing


- What is alternative splicing?
- What is possible to do computationally to better understand this complicated phenomenon?
  - Frequency of alternative splicing
  - Specialized databases
  - Search for regulatory elements

## PROCESSING mRNA








## The Nobel Prize in Physiology or Medicine 1993




The Nobel Assembly at the Karolinska Institute in Stockholm, Sweden, has awarded the Nobel Prize in Physiology or Medicine for 1993 jointly to Richard J. Roberts and Phillip A. Sharp for their discovery of split genes.

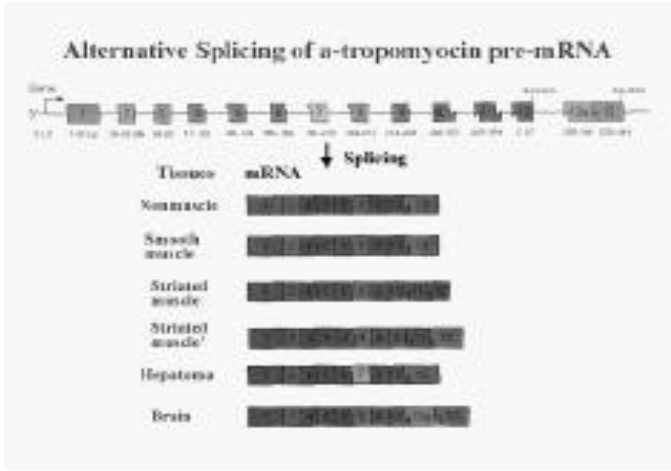
Computational Biology @ SC 2000



## α-Tropomyosin pre-mRNA

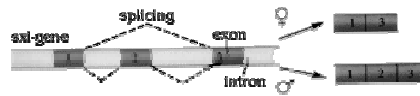


### Alternative Splicing of α-tropomyosin pre-mRNA



Computational Biology @ SC 2000

- A precursor-RNA may often be matured to mRNAs with alternative structures. An example where alternative splicing has a dramatic consequence is somatic sex determination in the fruit fly *Drosophila melanogaster*.



- In this system, the female-specific *xtl*-protein is a key regulator. It controls a cascade of alternative RNA splicing decisions that finally result in female flies.
- Sex in *Drosophila* is largely determined by alternative splicing



- Splicing errors cause thalassemia
- Thalassemia, a form of anemia common in the Mediterranean countries, is caused by errors in the splicing process.
- Normal red blood cells contain correctly spliced beta-globin, an important component in hemoglobin that takes up oxygen in the lungs.







## Information on alternative splicing in public databases:



- Swiss-Prot (protein) database is well curated, but the information content is incomplete with reference to alternative splicing and does not allow for automatic retrieval of such entries.
- Swiss-Prot entries just state the fact that a particular protein is one of the products of alternative splicing.
- Some entries contain the information on the limited number of isoforms.

Computational Biology @ SC 2000



## Clustering procedure



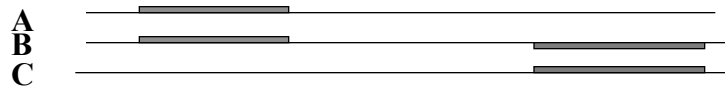
### Similarity analysis of two sequences

- |                                                                                                   |                                                                                              |
|---------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------|
| ■ Gene families<br>multiple similar genes<br>exist due to duplication<br>and divergence of genes. | ■ Alternative splicing<br>one gene but primary<br>transcript spliced in more<br>than one way |
| ↓                                                                                                 | ↓                                                                                            |
| ■ Short similar fragments,<br>a lot of mutations                                                  | ■ Relatively long identical<br>fragments                                                     |

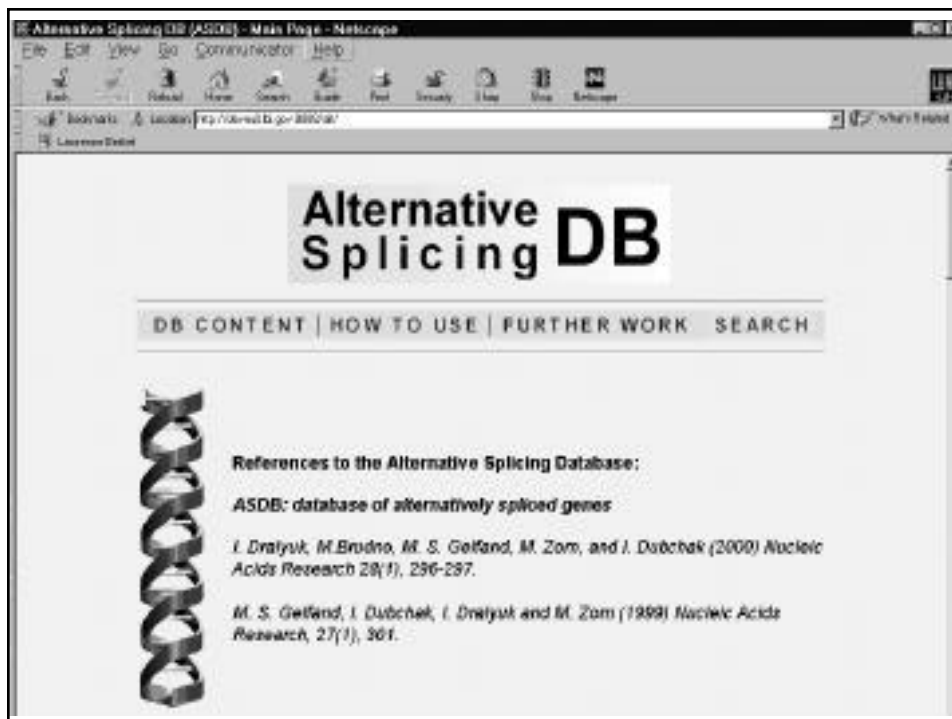
Computational Biology @ SC 2000

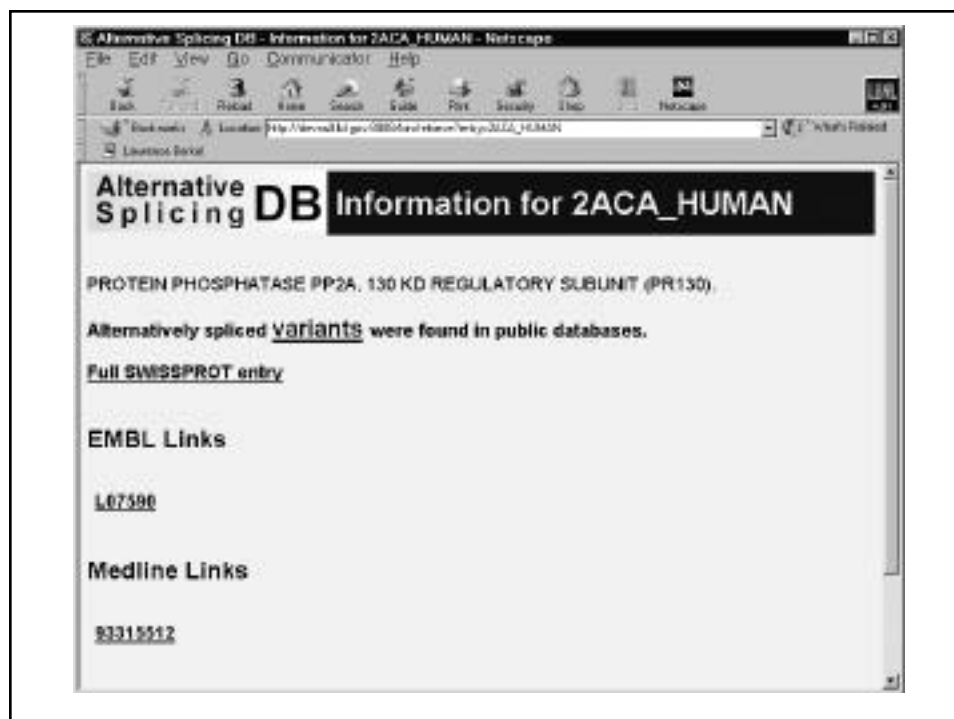
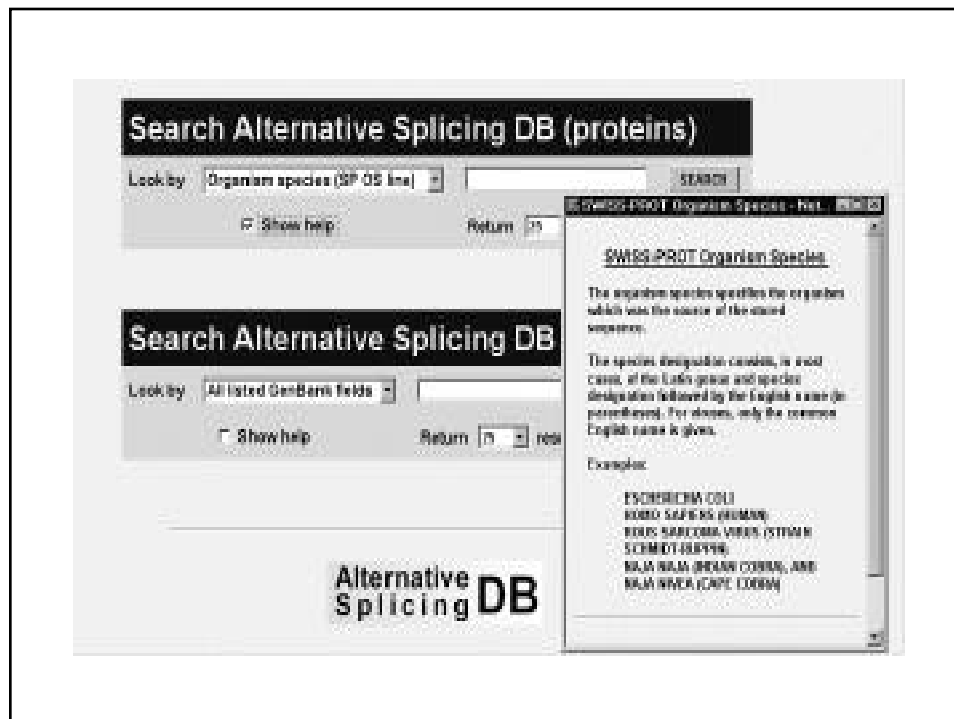
# Clustering procedure

- 1,922 protein sequences were compared all-against-all in order to find common sequence fragments.
- The length of this fragment was a variable parameter in the software. Various lengths were tested to cluster as many variants of the same gene as possible, but to avoid false clusters generated by too short fragments.



~ 240 clusters of isoforms





AS Alternative Splicing DB - Cluster Information - Netscape

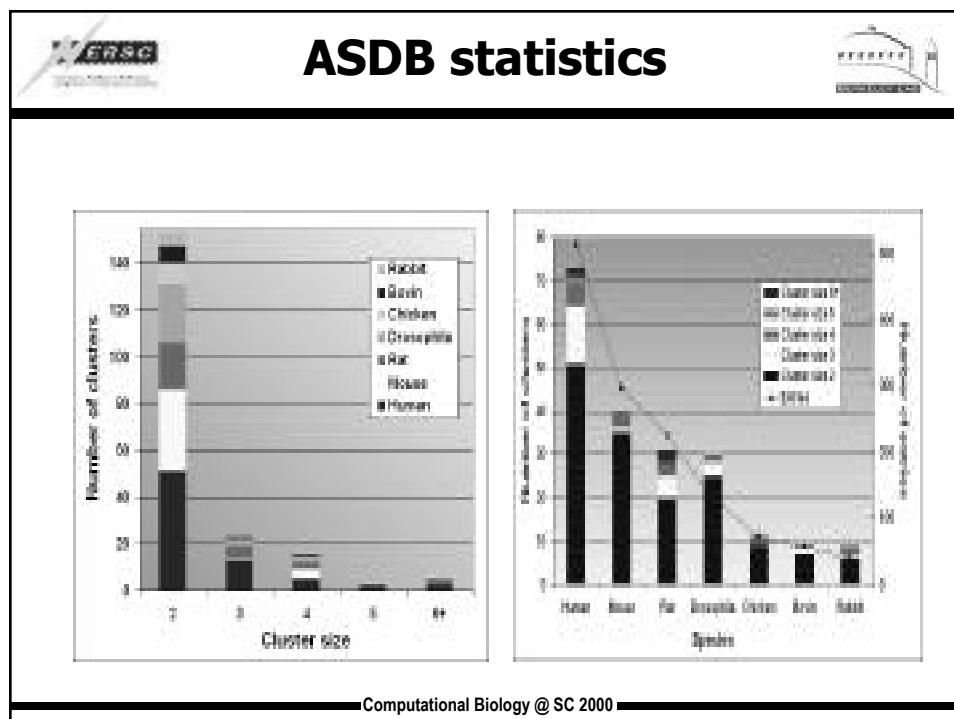
File Edit View Go Communicator Help

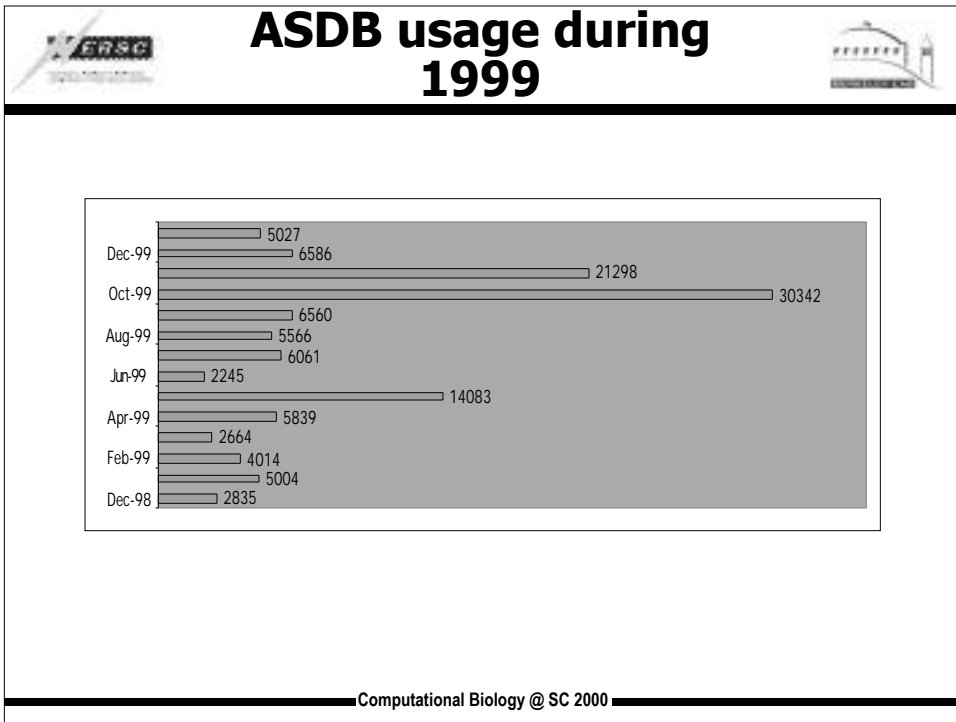
Back Forward Home Search Save Print Security Help

Location: <http://www.scripps.edu/ASDB/cluster/cluster.html>

Cluster: 2001

ZACA_HUMAN	IELQNDKFRD RRCDFVQSD NHPNSLYNI EVNDGDTLKA WQVGGGLIM
ZACD_HUMAN	.....
ZACA_HUMAN	NPLENVSSDD INRTIVIRK SDGERALDNG QTRSGGFRH LKRVNRRAR
ZACD_HUMAN	.....
ZACA_HUMAN	FTGHATHKE CPTNGGBIS KIPKSPFNI EMDCKSEVS KPEEGDQDF
ZACD_HUMAN	.....
ZACA_HUMAN	THSSGGBID KLMDLERFC QKHETLSTF LARGENMEL MHSGLTGOT
ZACD_HUMAN	.....
ZACA_HUMAN	LVDGERSKV SSVEKVSFS CLTRIETNG MHIEDRAL LKILGISID
ZACD_HUMAN	.....
ZACA_HUMAN	PAQELVCKE SSGSLQGRH MHQILQETL TSSQANLVC RSPVGDRAK
ZACD_HUMAN	.....
ZACA_HUMAN	PDRAVLIGQ PEVIRIQNEP SEKPGSLP DATSGDQRP LSPVWMMV
ZACD_HUMAN	PDGFKRLK PQQTQIQNEP SEKPGSLP DATSGDQRP LSPVWMMV
ZACA_HUMAN	WHALEINIS RYEEEGLED TCHREUTL RKTARMDS DQKADYENG
ZACD_HUMAN	WHALEINIS RPYFEGLED TCHREUTL RKTARMDS DQKADYENG
ZACA_HUMAN	KIAKVGCVL YKATFRMAA GGEKGFVA QSTIANKEEL LNNHDCASK
ZACD_HUMAN	KIAKVGCVL YKATFRMAA GGEKGFVA QSTIANKEEL LNNHDCASK
ZACA_HUMAN	PICLLAHYD SGLQGEDID ILQDVTGK GLTFLEADG FRERYTTVI





- Study of Regulation**
- No systematic surveys to address the relative importance of such elements in the regulation of alternative splicing.
  - It is unknown as to whether regulatory words occur more frequently adjacent to alternative exons than in the rest of the genome.
  - It is not clear whether these elements enhance splicing of only a limited set of exons, or have a more general role.
- Computational Biology @ SC 2000



## Alternative Splicing Regulation



- A number of genomic sequence regulatory elements have been identified outside of traditional splice sites.
- The concept of splicing "enhancers" and "silencers" that promote or inhibit splicing at neighboring splice sites is well established.
- Many alternative exons are probably regulated by a combination of silencers and enhancers.

Computational Biology @ SC 2000



## Data Collection



- Automated processing of GenBank/Medline
- Manual analysis of abstracts & articles
- Collecting the sample

Computational Biology @ SC 2000



## BiSyCLES Search Options



- **BiSyCLES searches in the two databases, then establishes which of the retrieved entries are linked**

- ✓ Medline: +“alternative splicing,” tissue, muscle, brain, neuro\*, heart, regul\*, enhancer, silencer
- ✓ Genbank: +“alternative splicing” +“complete CDS”

- **Results:**

- ✓ ~300 abstracts
- ✓ ~50 relevant papers

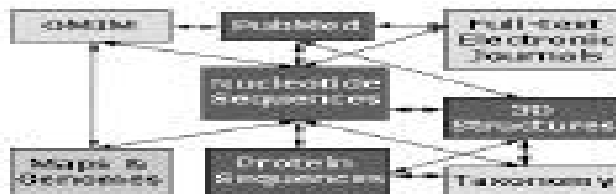
Computational Biology @ SC 2000



## BiSyCLES: Biological System for Cross-Linked Entry Search



- GenBank contains genomic data but little annotation
- Medline (PubMed) contains abstracts from journals but no genomic data
- NCBI's Entrez system keeps links between related entries in its databases



Computational Biology @ SC 2000



## Word Counting



- To calculate the confidence value of a particular word we select random subsets of a large dataset of constitutively spliced exons (1,504 exons; Burset & Guigo, 1996) equal in size to our alternative dataset.
- We then calculate the fraction of these subsets in which the word is over-represented at a higher rate than in the alternative set.
- (Over-representation is calculated as difference of frequencies)

Computational Biology @ SC 2000



## Known Regulatory Elements



<u>enhancers</u>	<u>reference</u>
UGCAUG	Huh & Hynes, 1994; Hedjran et al., 1997; Modafferi & Black, 1997; Kawamoto, 1996; Carlo et al., 1996
CUG repeat	Ryan et al., 1996; Philips et al., 1998
(A/U)GGG	Sirand-Pugnet et al., 1995a
GGGGCUG	Carlo et al., 1996
<u>silencers</u>	
UUCUCU	Chan & Black, 1995; Chan & Black, 1997; Ashiya & Grabowski, 1997

Computational Biology @ SC 2000





## Short summary



- In the simple cases of splicing, introns are always introns and exons are always exons
- During alternative splicing, within the same RNA, sequences can be recognized as either intron or exon under different conditions and the concept of exons and introns becomes rather empirical
- RNAs are not spliced differently in the same cell at the same time but in different cells or in the same cell types at different times in development or under different conditions
- A variety of patterns of alternate splicing have been observed.

Computational Biology @ SC 2000

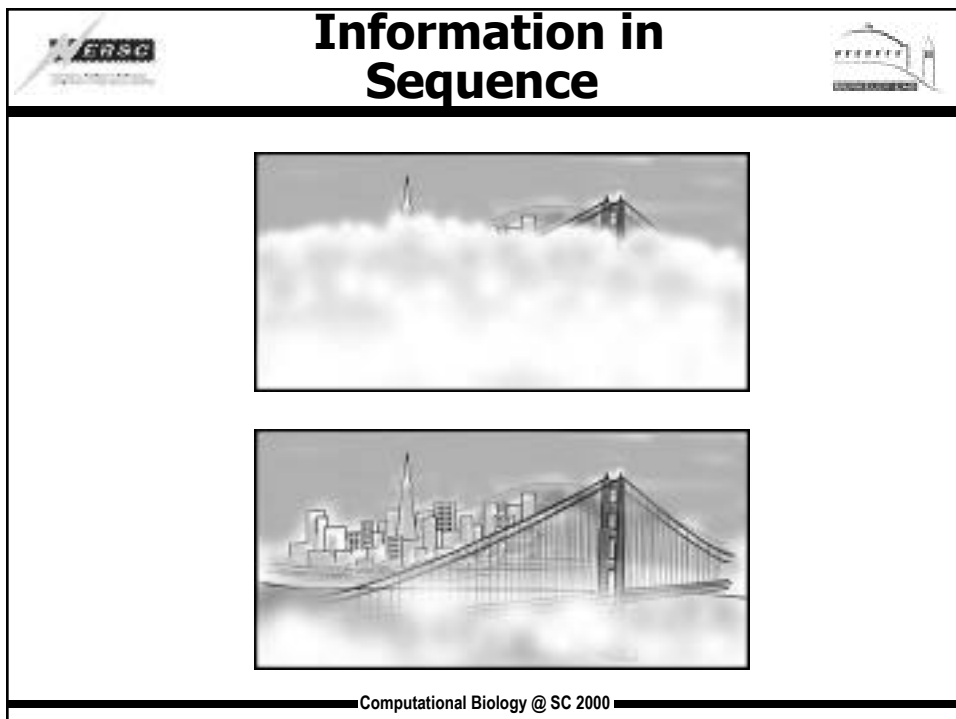
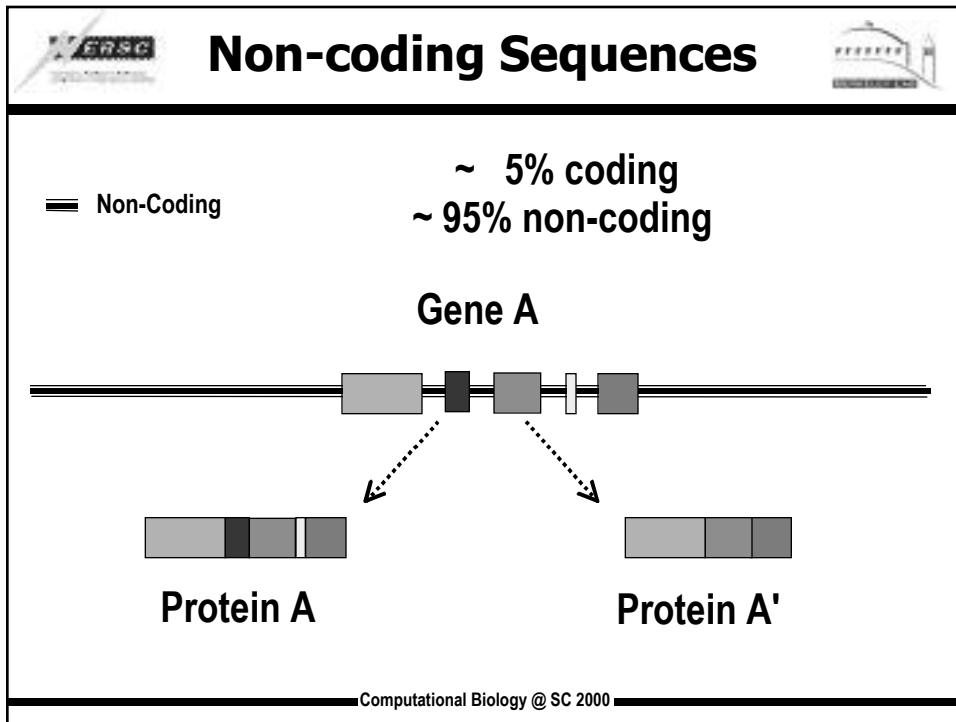


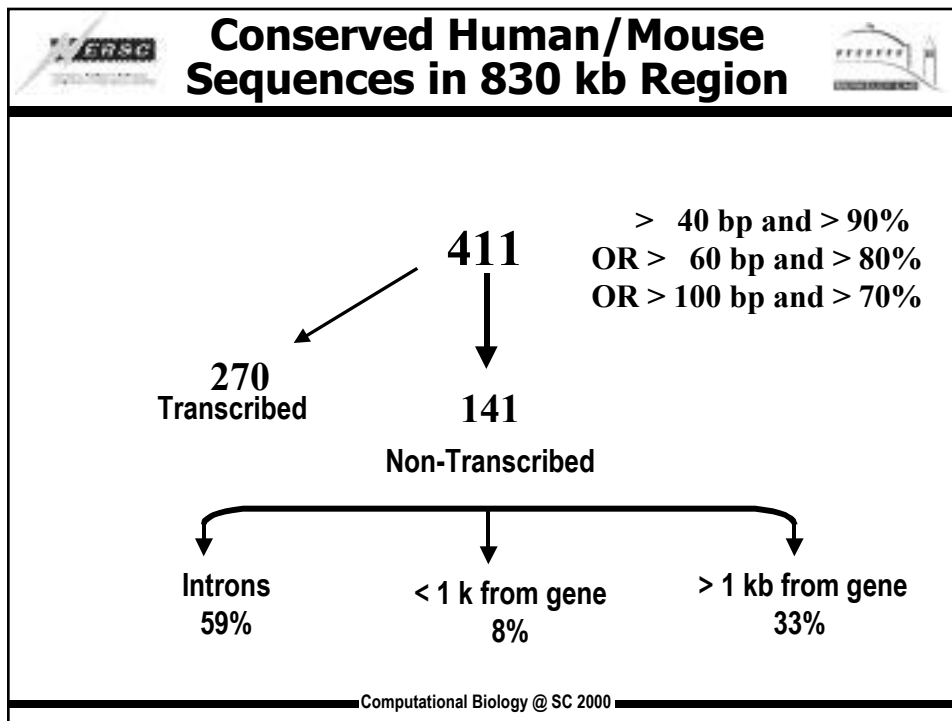
## Evolutionarily conserved non-coding DNA sequences



- Discovering them in DNA sequence
- Tools for their visualization
- Biological importance

Computational Biology @ SC 2000



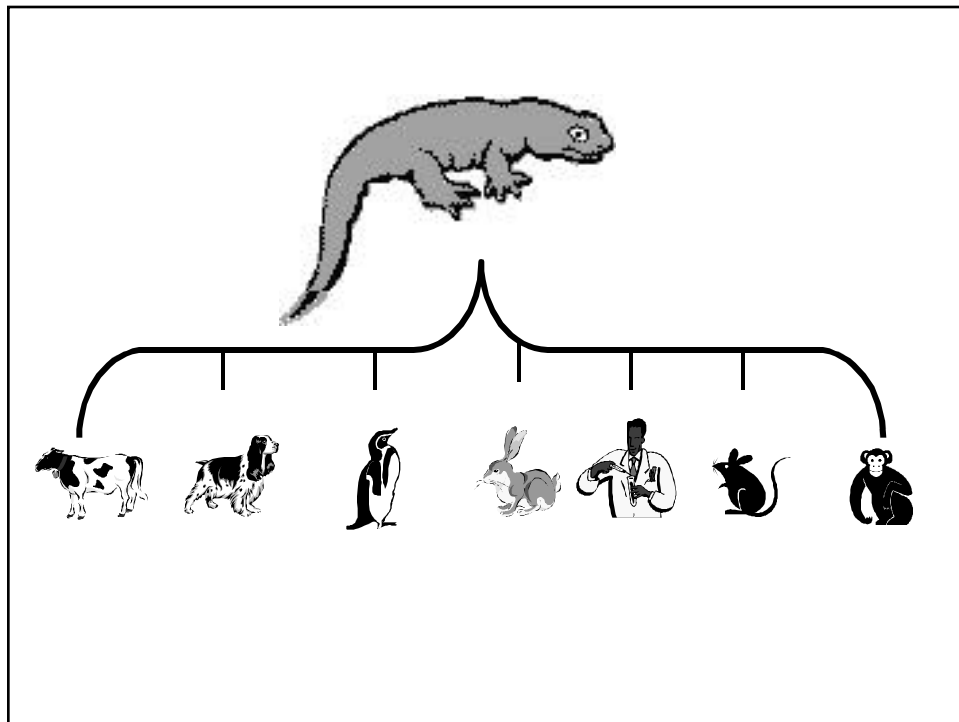




**90 Elements in 1 Megabase**

↓

**Are most conserved noncoding sequences "functional" or are they a product of passive evolution?**

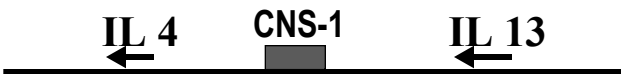
Computational Biology @ SC 2000





## Analysis of CNS-1

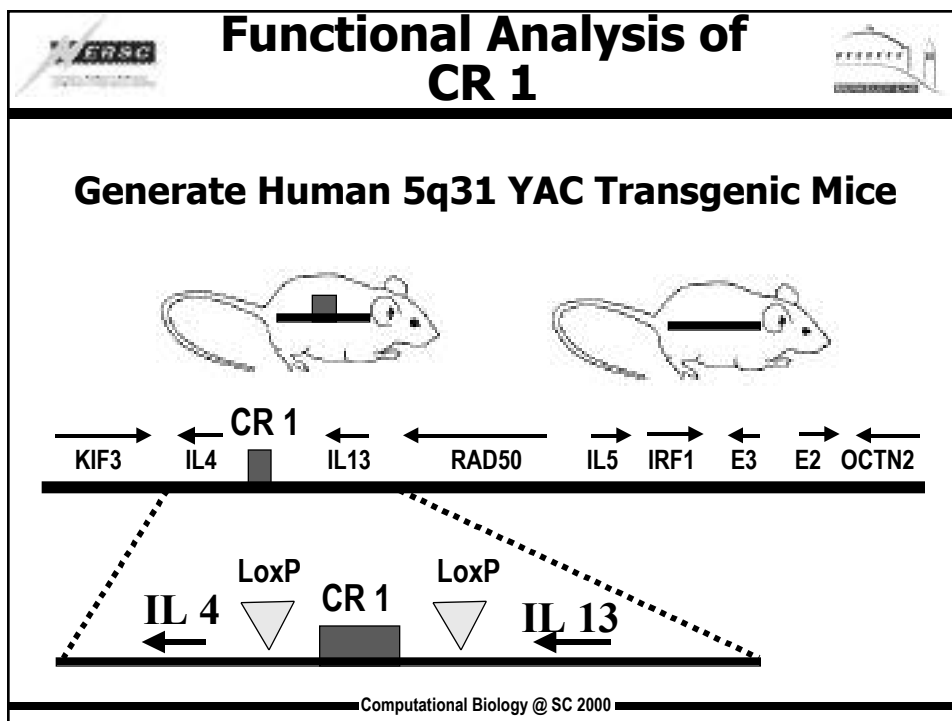
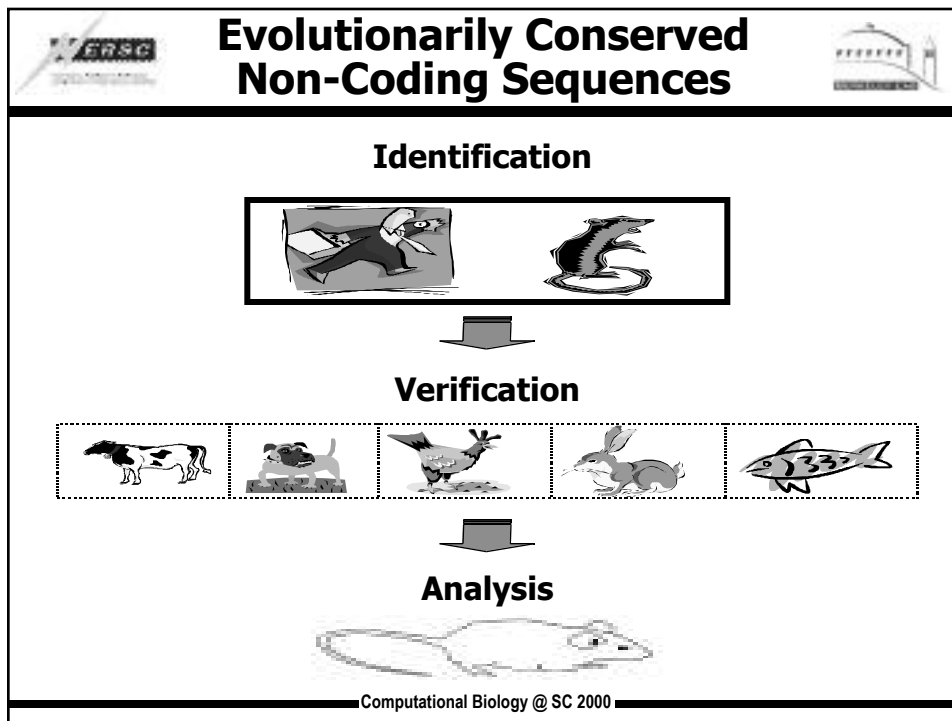
- Present in other species:
  - ◆ Cow (86%)
  - ◆ Dog (81%)
  - ◆ Rabbit (73%)
- Genomic position conserved in human, mouse, dog and baboon

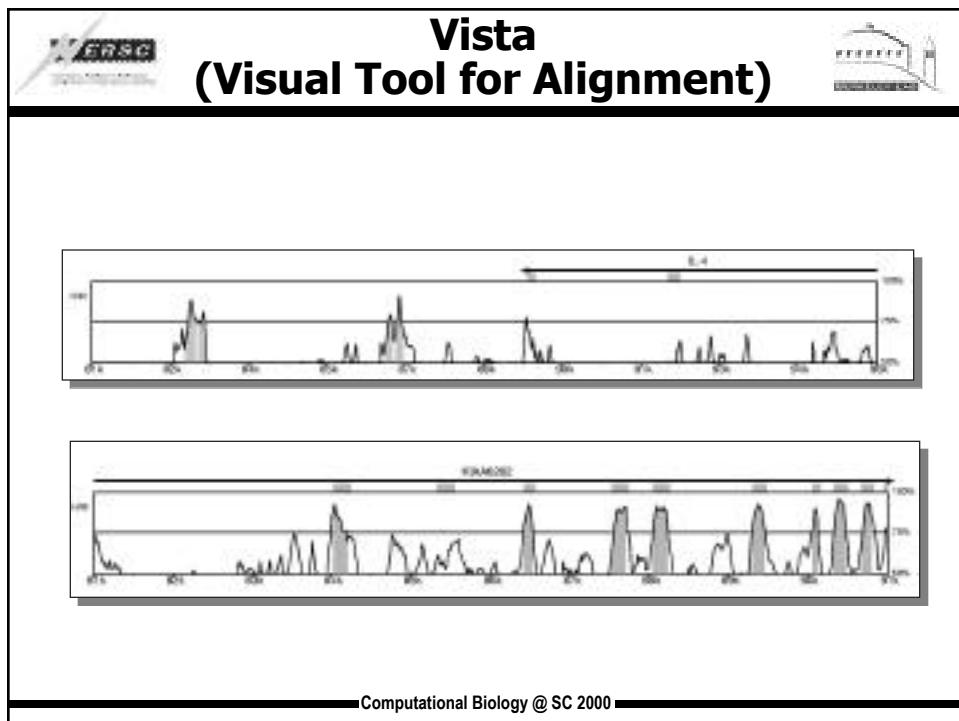
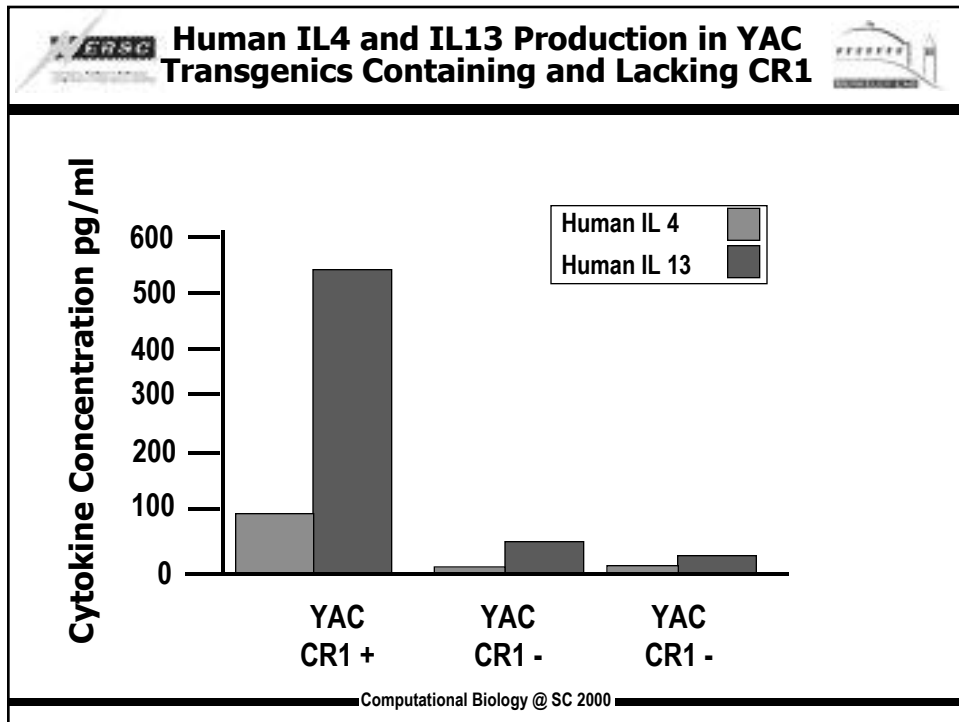


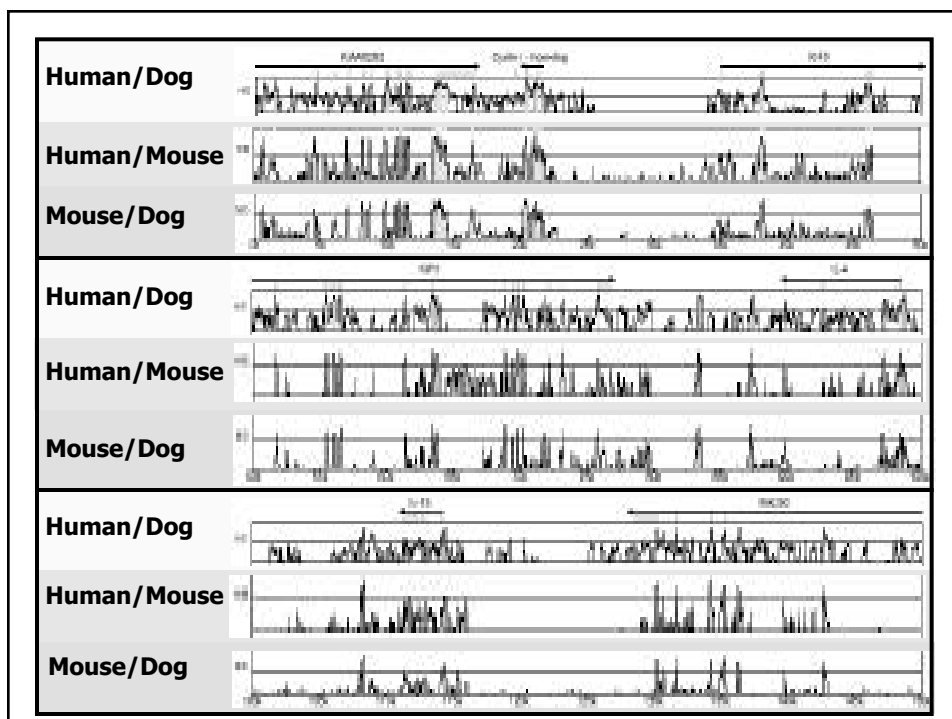
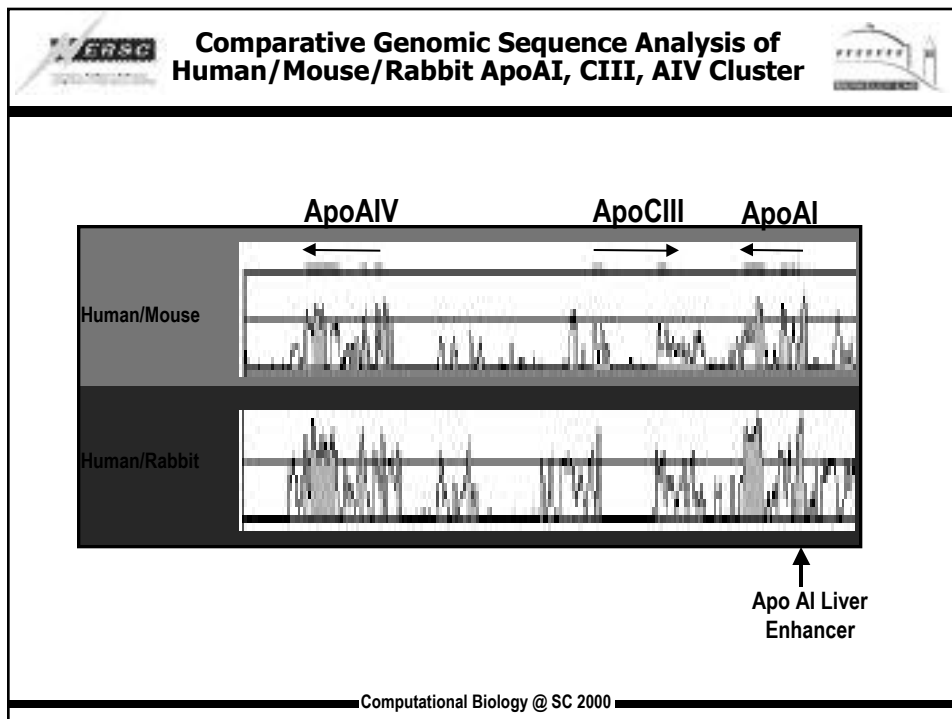
A horizontal line represents a chromosome. Above the line, from left to right, are the labels 'IL 4', 'CNS-1', and 'IL 13'. Below 'IL 4' and 'IL 13' are arrows pointing to the left. Below 'CNS-1' is a small grey rectangular box representing the gene's location.

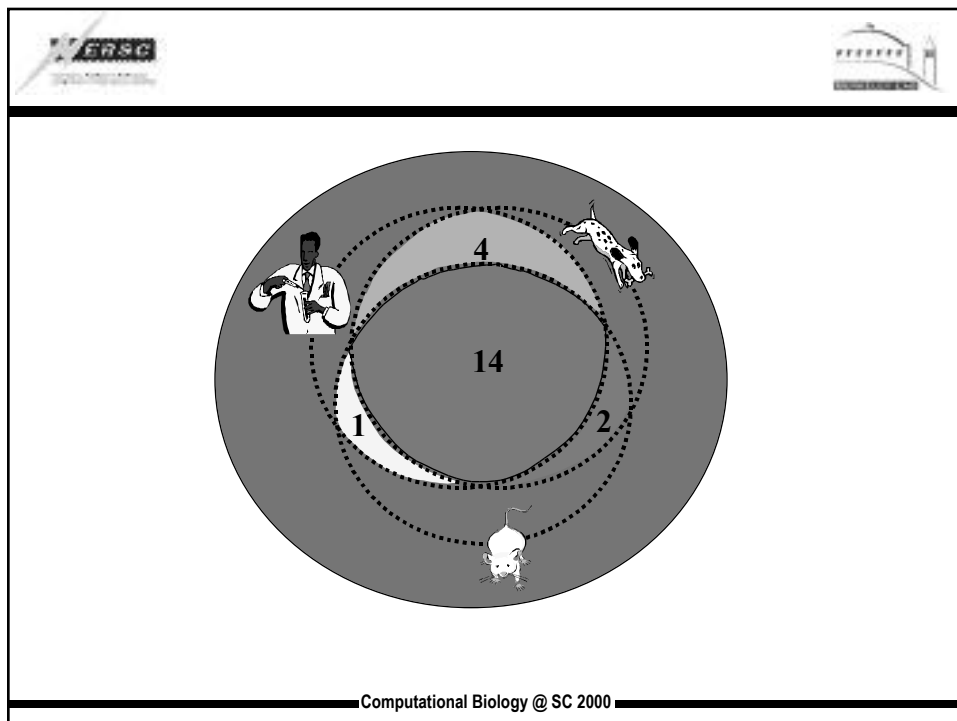
- Single copy in the human genome



Computational Biology @ SC 2000










 <http://www-gsd.lbl.gov/vista/> 



Welcome to the VISTA, or VISualization Tool for Alignments home page

VISTA is an integrated system for global alignment and visualization, designed for comparative genomic analysis:

1. The visual output is clean and simple, allowing the user to easily identify conserved regions.
2. Similarity scores are displayed for the entire sequence, thus allowing for the identification of shorter conserved regions, or regions with gaps.




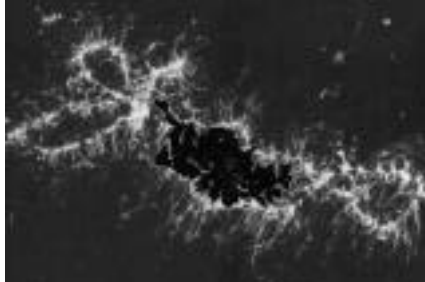
Computational Biology @ SC 2000

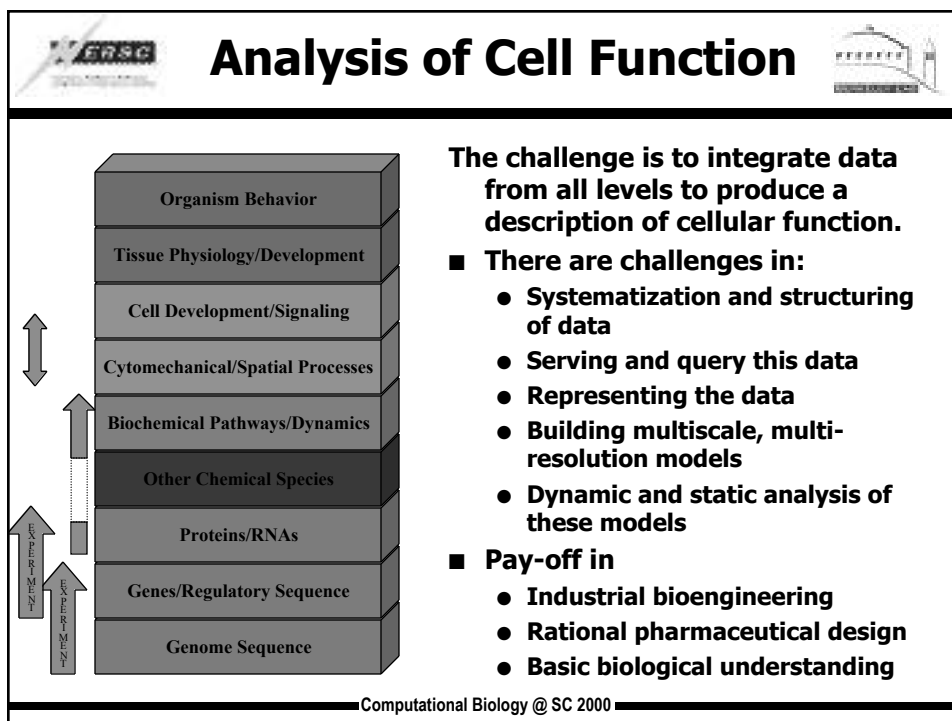
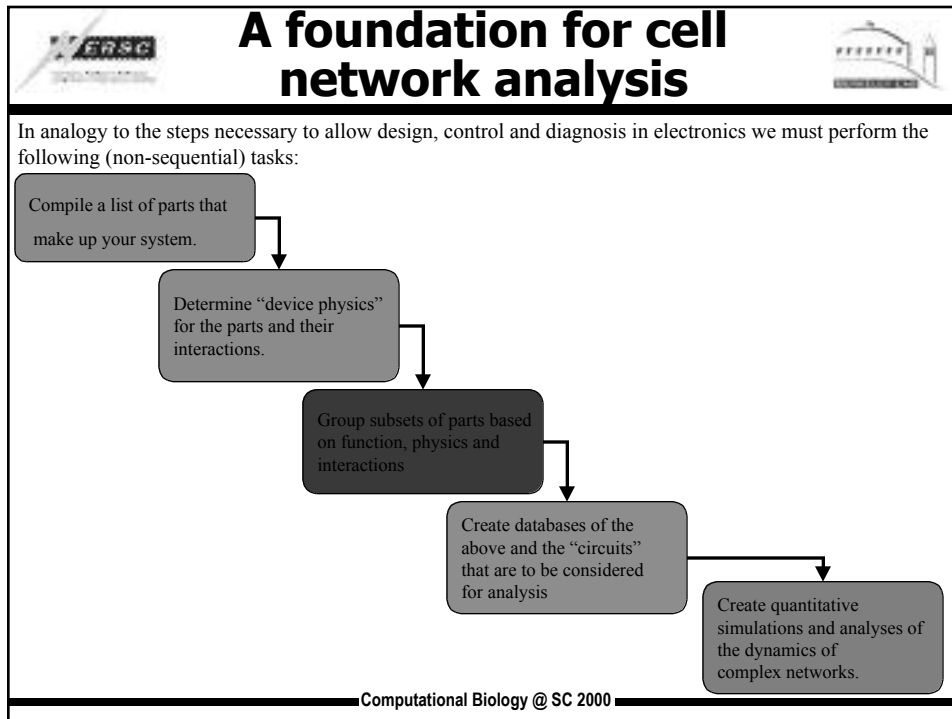


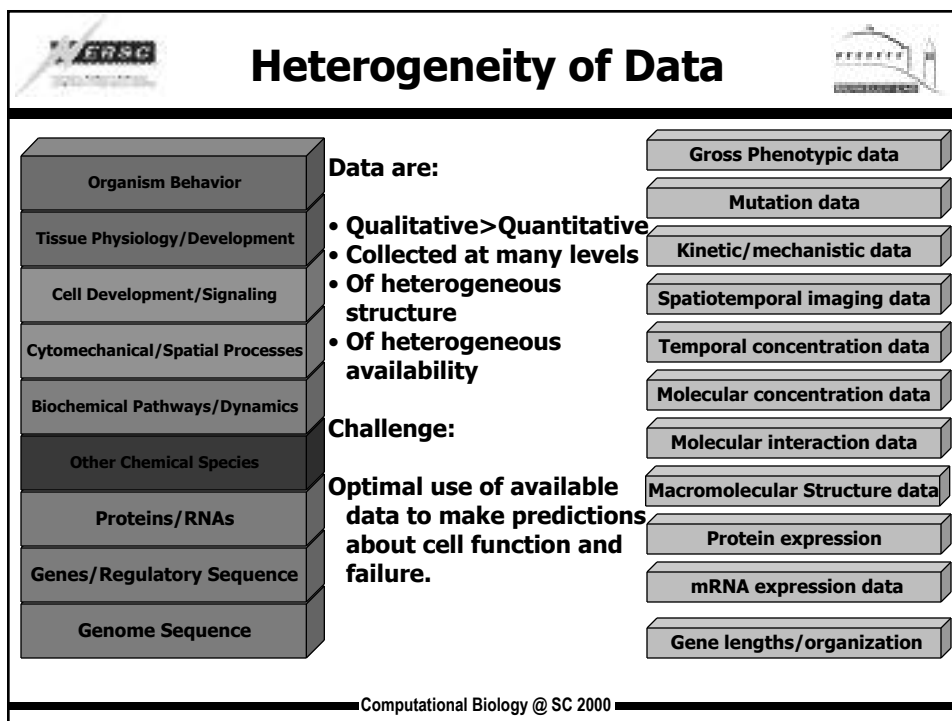
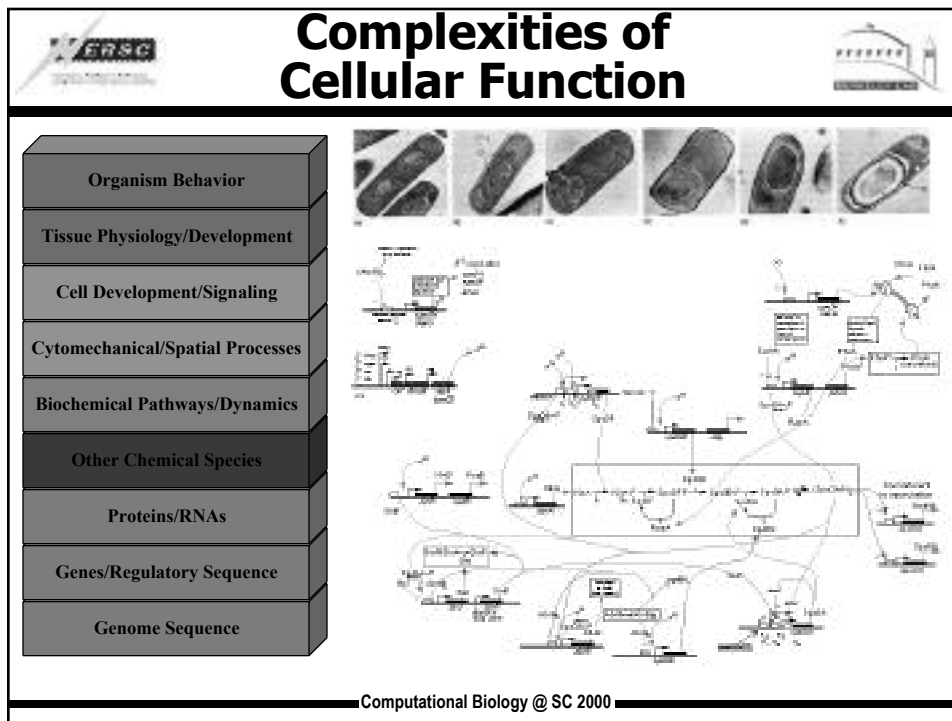


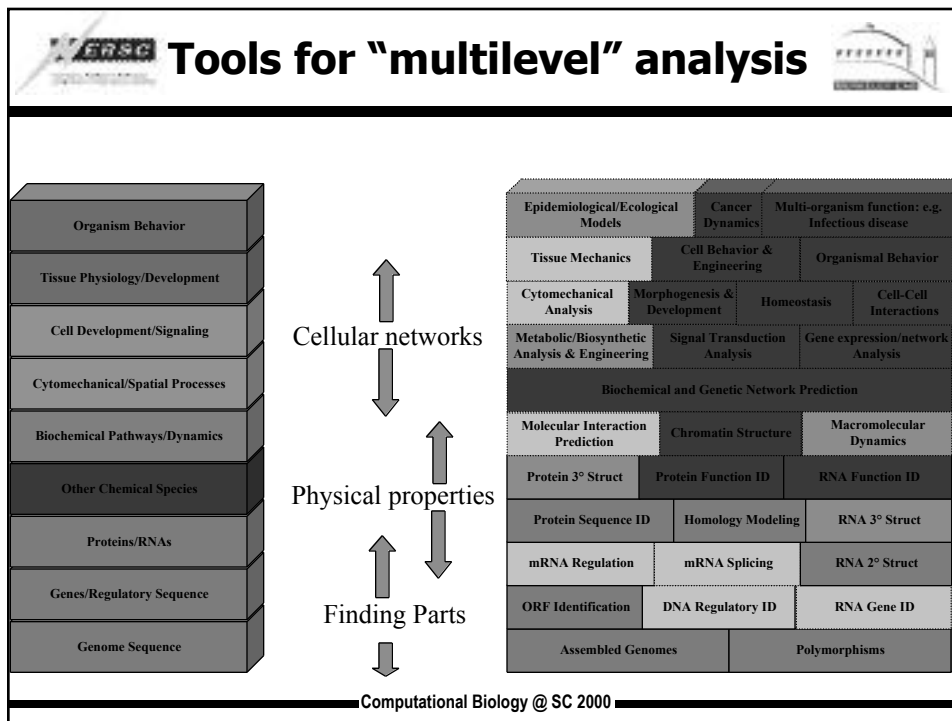
# Gene Regulatory Networks and Cellular Processes

Adam Arkin  
APArkin@lbl.gov  
LBNL

 <h2>Engineering of Cellular Circuitry</h2> 	
 <p>Courtesy of IBM</p> <p><u>Asynchronous Digital Telephone Switching Circuit</u></p> <p>Full knowledge of parts list Full knowledge of "device physics" Full knowledge of interactions</p> <p>No one fully understands how this circuit works!! Its just too complicated.</p> <p>Designed and prototyped on a computer (SPICE analysis) Experimental implementation fault tested on computer</p>	 <p>From: Wasserman Lab, Loyola</p> <p><u>Asynchronous Analog Biological Switching Circuit</u></p> <p>Partial knowledge of parts list Partial knowledge of "device physics" Partial knowledge of interactions</p> <p><b>No one fully understands how this circuit works!!</b> Its just too complicated.</p> <p>We need a SPICE-like analysis for biological systems</p>
<p>Computational Biology @ SC 2000</p>	









**Why now?**

- **Genome projects are providing a large (but partial) list of parts**
- **New measurement technologies are helping to identify further components, their interactions, and timings**
  - ✓ Gene microarrays
  - ✓ Two-Hybrid library screens
  - ✓ High-throughput capillary electrophoresis arrays for DNA, proteins and metabolites
  - ✓ Fluorescent confocal imaging of live biological specimens
  - ✓ High-throughput protein structure determination
- **Data is being compiled, systematized, and served at an unprecedented rate**
  - ✓ Growth of GenBank and PDB > polynomial
  - ✓ Proliferation of databases of everything from sequence to confocal images to literature
- **The tools for analyzing these various sorts of data are also multiplying at an astounding rate**

Computational Biology @ SC 2000



## SPICE Tools for Biology?

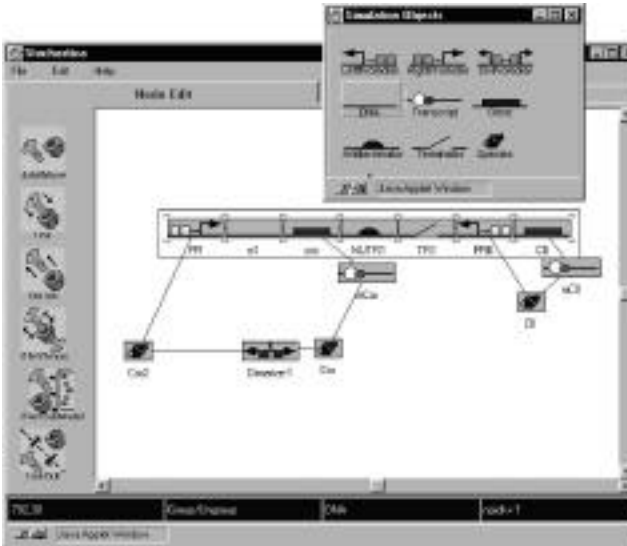


**Bio/Spice:** A Web-Servable, Biologist-Friendly, database, analysis and simulation interface was developed into a true beta product.

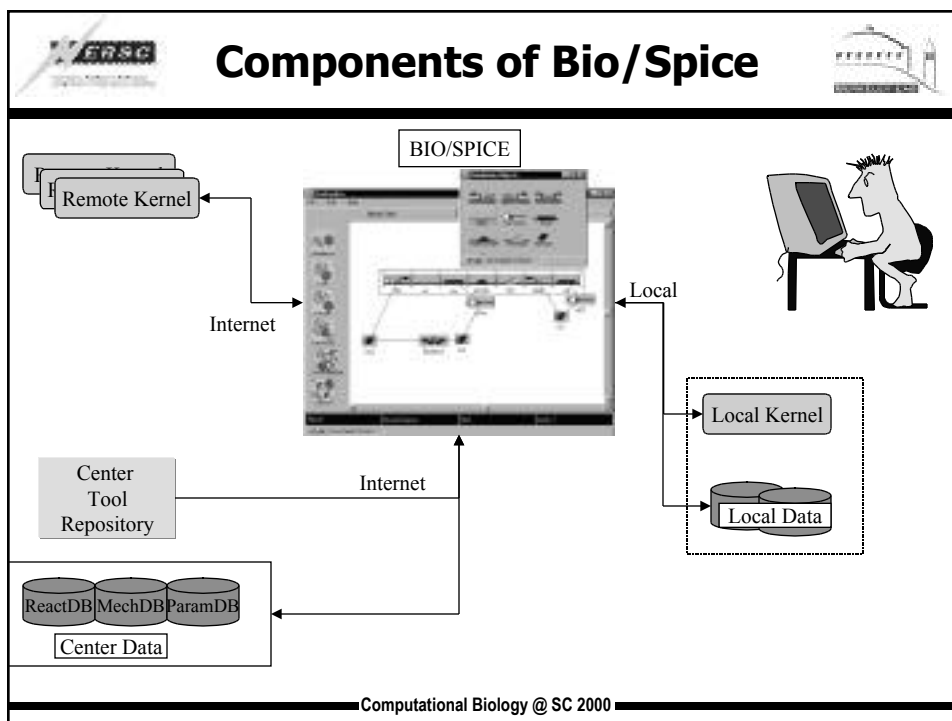
Interfaces to ReactDB, MechDB, and ParamDB.

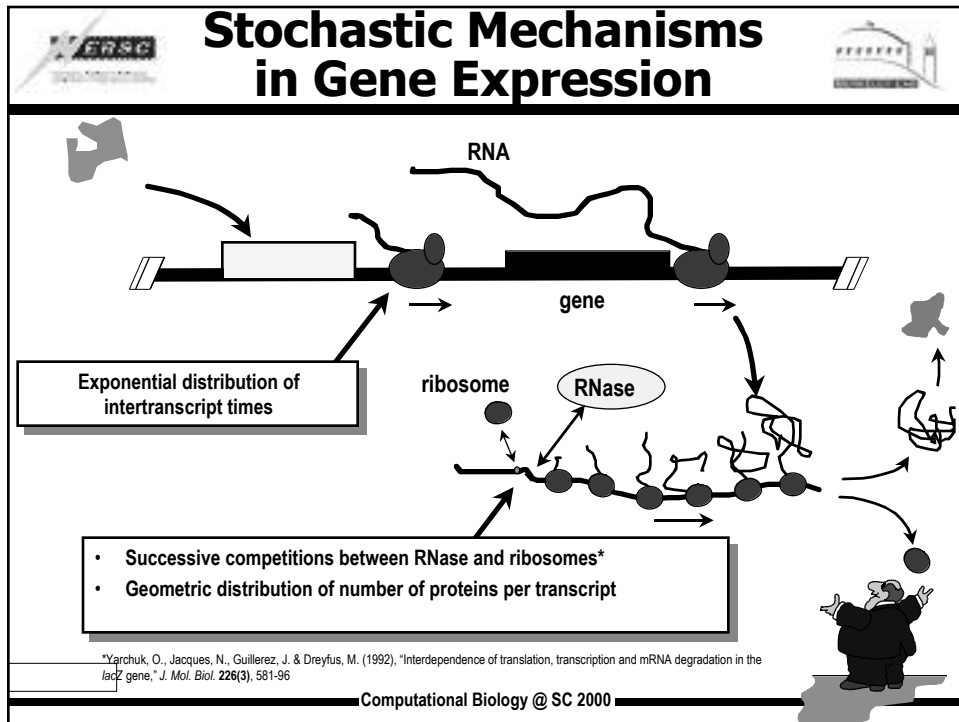
With Kernel, performs basic: flux-balance analysis, stochastic and deterministic kinetics, Scientific Visualization of results.

Notebook/Kernel design optimized for distributed computing.



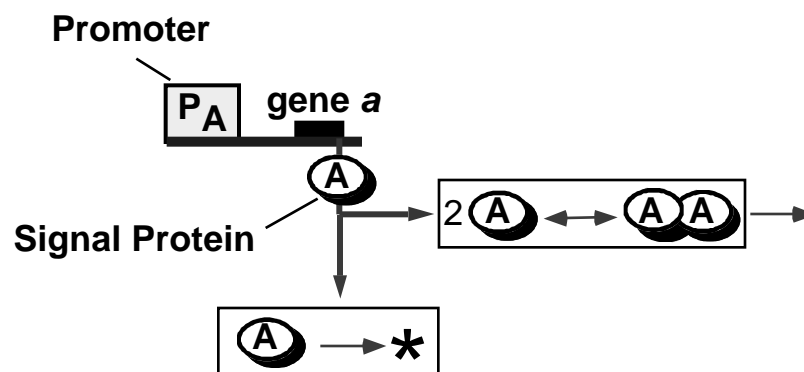
Computational Biology @ SC 2000

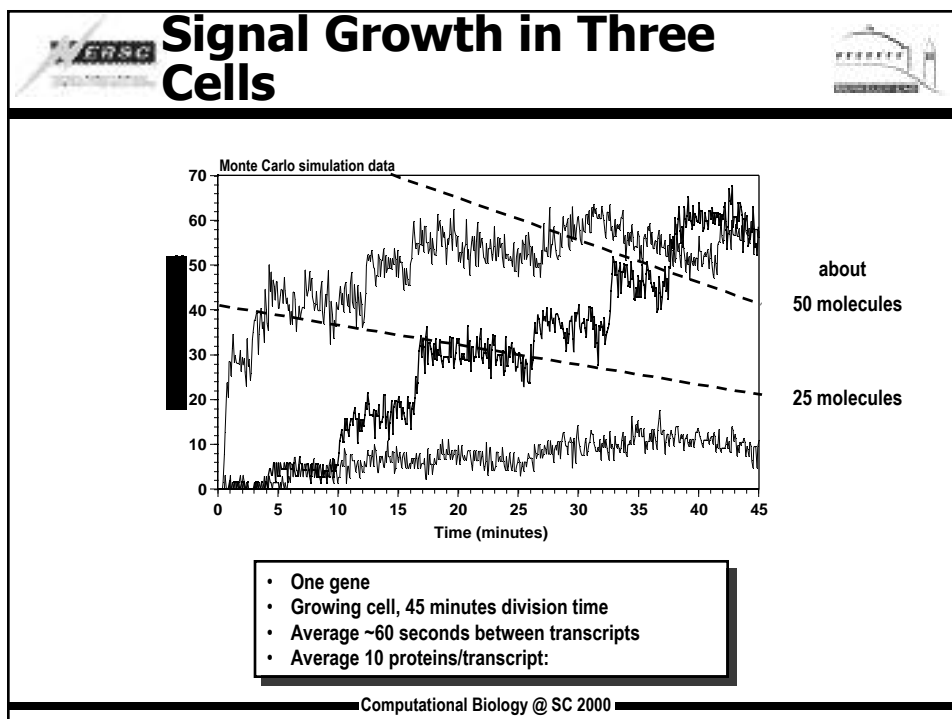
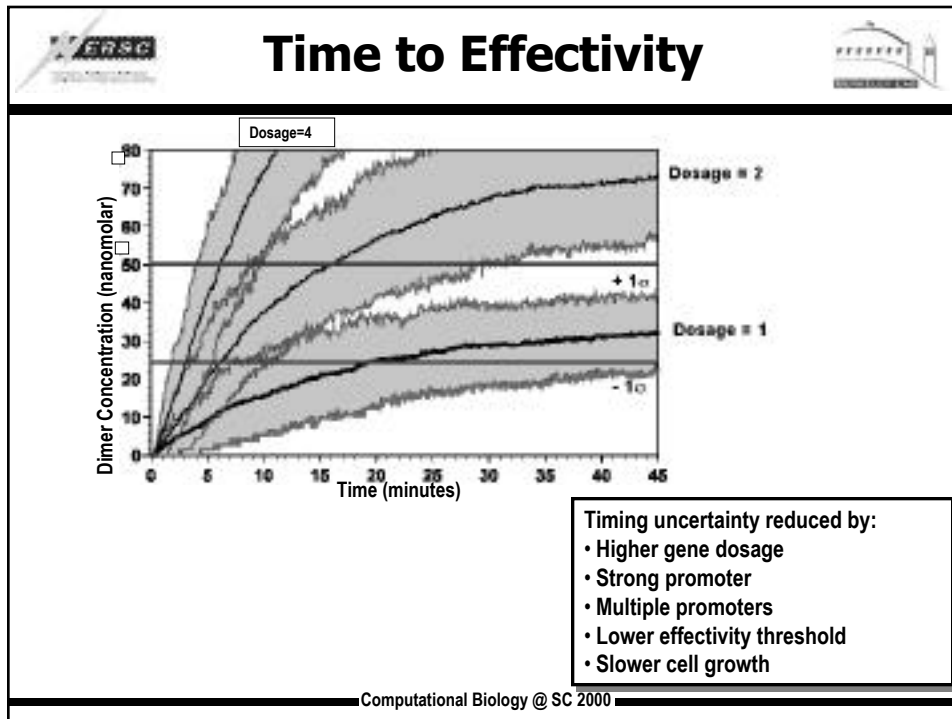




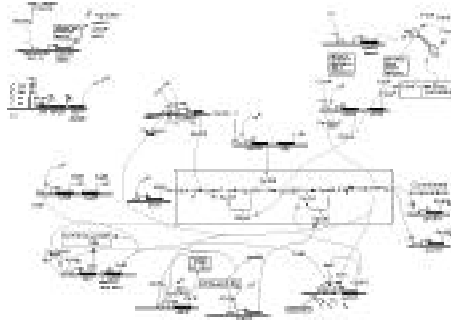
- # Some Stochastic Cellular Phenomena
- Lineage commitment in human hemopoiesis
  - Random, bimodal eukaryotic gene transcription in
    - Activated T cells
    - Steroid hormone activation of mouse mammary tumor virus
    - HIV-1 virus
  - Clonal variation in:
    - Bacterial chemotactic responses
    - Cell cycle timing
  - *E. coli* type-1 pili expression
    - Enhances virulence
  - Changing cell surface protein expression
    - For immune response avoidance
  - Bacteriophage  $\lambda$  lysis/lysogeny decision
- Computational Biology @ SC 2000

- Random environmental influences
- Mutations
- Asymmetric partitioning at cell division
- Stochastic mechanisms in gene expression
  - Stochastic timing of gene expression
  - Random variation in time for signal propagation
  - Random variation total protein production









This is approximately 1/3 of just the initiation of the sporulation program from *Bacillus subtilis*.

There are over 100 proteins, 40 genes, 300 reactions for which data is available.

The total data on just this process is a tens of Gb and it is incomplete. Microarray and microscope data are added 100 Mb per week. Model builders need to query this data and arrange it for simulation. Simulations must be run under many different condition and hypotheses.

## ■ Data Handling:

The total data necessary for network analysis is huge. By nature it will be distributed and heterogeneous

We need:

- ✓ Database standard and new query types
- ✓ Means of secure, fast transmission of information
- ✓ Means of quality control on data input

## ■ Tool integration:

- ✓ Centralization of computational biology tools and standards
- ✓ Ability to use tools together to generate good network hypotheses
- ✓ Good quality ratings on Tool outputs

## ■ Advanced Simulation Tools:

- ✓ Fast, distributed algorithms for dynamical simulation
- ✓ Mixed mode systems (differential, Markov, algebraic, logical)
- ✓ Spatially distributed systems

